UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title

RNA sequencing applications of a reverse transcriptase from non-LTR R2 retroelements

Permalink

https://escholarship.org/uc/item/5p02k9ns

Author

Ferguson, Lucas John

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

RNA Sequencing Applications of a Reverse Transcriptase from Non-LTR R2 Retroelements

By

Lucas John Ferguson

A dissertation submitted in satisfaction of the

requirement for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Kathleen Collins, co-chair

Professor Nicholas T. Ingolia, co-chair

Professor Britt A. Glaunsinger

Professor Donald C. Rio

Spring 2024

Abstract

RNA Sequencing Applications of a Reverse

Transcriptase from Non-LTR R2 Retroelements

By

Lucas John Ferguson

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Kathleen Collins, co-chair

Professor Nicholas Thomas Ingolia, co-chair

While high-throughput sequencing of messenger (m)RNA has become routine, analogous measurements of small and non-coding RNA (ncRNA) populations are challenging, limiting our view of the transcriptome. Likewise, sequencing-based measurement of protein synthesis by ribosome profiling have been possible for over a decade, but challenging workflows have limited its wide-spread adoption beyond specialists. From a technical standpoint, both endogenous small RNAs and ribosome-protected mRNA footprints (RPFs) lack well-defined 3′ sequences, such as polyadenylated tails, and thus require multiple enzymatic steps to append sequencing adapters before reverse transcription. These multi-step protocols have variable and inconsistent yields that introduce stochasticity, along with bias caused by preferential capture of enzymatically favorable substrates. Moreover, existing small RNA library generation approaches are notably labor intensive, unsuitable for automation, and require high RNA inputs, limiting the scale and scope of their use. In this dissertation, I created a new approach that removed these barriers, and I used it to uncover new biological insights.

To address the technological shortcomings of small and ncRNA profiling, I co-invented, optimized, and applied a novel approach for small RNA library generation: Ordered Two-Template Relay (OTTR). This method takes advantage of a recombinantly expressed reverse transcriptase (RT) derived from a non-long terminal repeat R2 retroelement in Bombyx mori to seamlessly convert small RNA into cDNA with flanking sequencing adaptors in a single-tube, 4-hour, workflow. Critically, this dramatic simplification does not compromise the quality of the libraries; instead, it notably reduces library generation bias across a range of substrates, such as micro(mi) RNAs, RPFs, and transfer(t) RNAs and tRNA-derived fragments. In this dissertation, I identified potential sources of bias in OTTR and modified the protocol to mitigate this residual

bias. These optimized conditions have since been made commercially available as of the completion of this dissertation, leading to the distribution and sale of OTTR cDNA library synthesis kits.

Beyond the invention and optimization of OTTR, I also completely redeveloped ribosome profiling, a technically demanding and bias-sensitive technique for quantifying transcriptome-wide ribosome occupancy through sequencing of RPFs generated by nuclease digestion. I found OTTR reduced ribosome profiling library bias compared to the traditional ligation-based protocol, leading to the adoption of OTTR for ribosome profiling. Furthermore, I established P1 nuclease as a valuable tool for RPF generation independently of OTTR. I found that RPF production by P1 nuclease was less disruptive for RPF-ribosome interactions and limited ribosomal RNA (rRNA) cleavage, greatly improving the yield of RPF sequencing reads by reducing rRNA contaminants. These improvements also reveal a wide array of protected mRNA fragments resulting from several types of ribosome-ribosome collisions, which are increasingly recognized to offer insight about translation efficiency and nascent protein folding.

Ribosome-ribosome collisions have been observed in both eukaryotic and prokaryotic cells, and specialized pathways to resolve these collisions and destroy problematic mRNA are under intensive study for their role in protein quality control. In recent years, a connection between ribosome traffic flow and cellular stress has been uncovered. With collaborators, I applied ribosome profiling with P1 nuclease and OTTR to uncover an unappreciated connection between selenium deficiency and ribosome stalling on selenocysteine-encoding mRNAs linked to LRP8, a ferroptosis resistance factor upregulated in cancer cells. Beyond ribosome pausing, I was able to synthesize and study ribosome profiling libraires made to investigate different questions, such as how a SARS-CoV-2 protein remodeled the transcriptomic and translatomic landscape of a cell, how a ribosome profile changes during neuronal progenitor cell differentiation, or how T-cell activation poises certain mRNAs for robust translation.

Overall, OTTR enabled rapid and efficient cDNA library synthesis across a range of small RNAs, and is well suited for low-input samples, machine automatization, and can even be extended for sequencing of highly fragmented and damaged DNA. This method simultaneously closes the gap between transcriptomic mRNA profiling and small and ncRNA profiling, while opening the door for high-throughput studies on ncRNA abundance and processing patterns in normal and diseased cells.

Dedication

To all of those who gave their youth to the pursuit of education, their careers to the waves of mentorship, and their lives to the march of science—and to those who support and love scientists, waiting patiently for their loved ones in the lab—I dedicate this dissertation to you.

"[…] and everything runs by Sanger's rule; that anytime you get technical development that's two to threefold or more efficient, accurate, cheaper, a whole range of experiments opens up."

David Schlessinger, Ph.D.
Scientist Emeritus
Laboratory of Genetics and Genomics
National Institute on Aging

Table of contents

Abstract 1
Dedication i
Table of Contents ii
Acknowledgements iv
Chapter 1: Biochemistry of non-LTR R2 reverse transcriptases and OTTR 1
Biochemistry of non-LTR R2 reverse transcriptases. 2
The invention of OTTR 5
Chapter 2: Improved precision, sensitivity, and adaptability of Ordered Two-Template Relay cDNA library preparation for RNA sequencing 9
Abstract 9
Introduction 9
Results and Discussion 13
Materials and methods 29
Tables 35
Chapter 3: Streamlined and sensitive mono- and di-ribosome profiling in yeast and human cells 42
Abstract 42
Introduction 42
Results 44
Discussion 75
Materials and methods 76
Tables 86
Chapter 4: Additional considerations and future improvements to OTTR 87
References 90

Acknowledgements

At the threshold of graduation, I find myself reflecting on the extraordinary individuals who have made my journey not only possible but profoundly enriching. Overwhelmingly, it was the educators in my life who have had the greatest influence on my direction. To all faculty and staff of the Mississippi School for Mathematics and Science, past and present—a deep and profound thank you for the immeasurable impact you have had on my life. To the late Executive Director, Dr. Charles Brown—thank you for believing in me and giving me the second chance I needed. To Dr. Henry Xiu-Feng Wan—thank you for opening the very first lab doors to a young and eager Lucas and shepherding my research throughout my undergraduate career. To Dr. Bill Epperson, Dr. Alicia Olivier, and Dr. Amelia Woolums—thank you for introducing me to the world of veterinary medicine, epidemiology, and disease pathology. To Dr. Thomas Anderson—thank you for believing in me and encouraging me to apply and become a Gates-Cambridge Scholar. To Dr. Andrew Firth, Dr. Ian Briarly, and Dr. Valeria Lulla—thank you for supervising my Cambridge research, which established a strong foundation for my dissertation research.

To my co-mentor Professor Nicholas Ingolia—thank you for graciously guiding my bioinformatic pursuits toward fruitful trees and kindly nudging me away from distracting flowers. To my co-mentor and fellow co-inventor Professor Kathleen Collins—thank you for devoting your life to the lab and establishing the most nurturing, warm, kind, and productive space I have ever had the privilege of working. Kathy and Nick, your impact on my life and words of advice will echo in my mind for decades. I am so lucky to have had your mentorship, and I will forever miss our weekly Trio meetings. Over the years, I have grown to deeply respect and care for the efforts you put into each graduate student you mentor. I realize how lucky I am and will spend the rest of my life trying to mentor others as well as you have mentored me—all I can say is thank you.

To Dr. Heather Upton and Ms. Sydney Pimentel—thank you for co-inventing Ordered Two-Template Relay and encouraging me to keep moving forward with our invention. Working together as a team towards a common goal was a unique graduate experience that has influenced my career trajectory and goals.

Pursing this degree as a co-mentored student gave me two different communities to belong to: the Collins Lab and the Ingolia Lab. To the Collins Lab, thank you for being incredible comrades on the “dye-front” of nucleic acid and polymerase biochemistry. Learning together, perfecting our techniques, and troubleshooting our experiments has been among the sweetest moments during the pursuit of this degree. Each opportunity to brainstorm new experiments, or commiserate failed ones, reminded me how far each of us have come in our pursuit of knowledge. I have cherished being able to contribute in my own way to your pursuits, and I hope each of you continue doing amazing work and keep me updated throughout your careers.

To Dr. Briana Van Treeck, thank you for being so kind and reassuring under all circumstances. You will always be a role model to me as I become a mentor to others.

To Anthony Rodríguez-Vargas and Jeremy McIntyre, thank you for joining all those years ago so I was no longer the only Collins Lab graduate student, and for all the years of comradery since—from the COVID-19 pandemic to the final chapters of our degrees, it has been an honor to have overlapped these years with you two. To Connor Horton thank you for being so incredibly funny and always willing to share a laugh—please keep my computer monitor as a token of my sincere appreciation. To Justin Shumate thank you for being a genuine community builder and always striving to make the Collins Lab an environment where anyone can feel welcomed. To Kat Flug, Sarah Palm, and Anastasiya Trzcinski, thank you for reminding me the sense of wonder and excitement that comes from learning a new field—I hope to channel this as I embark on my own new fields of study. I wish you three the most successful of graduate careers, and I cannot wait to see what comes next.

To past members Rosa Lee, Sydney Pimentel, Dr. Heather Upton, and Xiaozhu Zhang, thank you for each of your indelible marks on the lab and my life. As members leave their impacts linger on in the lab for years to come, through spoken word between colleagues or notes in protocols. Soon, I will leave lab and my name will linger on as yours has. Together, our names will pass through the mouths of future Collins Labbites until smoothed over by passing time—I welcome the tide.

To the Ingolia Lab, and the Lareau Lab, thank you for providing an enriching space to strengthen my skillsets in both quantitative biology and scientific communication. Being able to share my co-mentored research across three distinct lab groups throughout my degree has been instrumental in the development of my presentation skills—thank you for the years of feedback and the excellent environment to practice and hone these skills. I would like to thank Dr. Ryan Muller for mentoring me during my first rotation, and for teaching me many valuable techniques I still rely on today. To Dr. Sam Fernandez and Dr. Paige Diamond, thank you for the years of scientific discussion and advice throughout my degree. To post-doctorates Dr. Joe Lobel and Dr. Wren Kim, thank you for sharing your constructive feedback throughout the years, and all your advice and thought-provoking questions. To Anna Freitas and Jinyoung Kim, you two are officially the senior Ingolia Lab graduate students now! As you move forward in your degrees do not hesitate to seek my advice when needed -I am just an email away. And to my fellow colleagues on Compute 1 – 4, thank you for sharing these computational resources with me and being virtual companions on the server. It has been a BLAST!

Out of a profound sense of respect, I would also like to express gratitude to the resources, facilities, and fellow employees of the University of California, Berkeley, which made my work comfortable, clean, safe, productive, and possible. To my co-mentor Prof. Nicholas Ingolia, thank you for all your efforts in maintaining and administering Compute 1 – 4, without these computational resources my work would have been severely limited. To the past and current staff of the Graduate Affairs Office, thank you for all which you do to ensure each graduate student in our department is supported and moving towards the finish line. To the scientists, staff, and technicians of the Vincent J. Coates Genomics Sequencing Lab, UC Berkeley DNA Sequencing

Facility, Electron Microscope Lab, and Macrolab—thank you for all that you do to support on-campus research scientist like me. To the Barker Hall Storeroom, thank you for all that you do to make research in Barker Hall as convenient and hassle-free as possible. To Ms. Loida Pineda, thank you for your diligent and thorough efforts in keeping the Collins Lab stocked and sterile. To all the staff, visible and invisible, who maintain, clean, and otherwise keep Barker Hall operational—a sincere and heartfelt thank you.

Of course, I would also like to acknowledge those who nurtured me towards being a better human. To my mother, Julie Bost Ferguson, and my father Travis Gwynn Ferguson—thank you for providing me everything, thank you for giving me the agency to choose my own path, thank you for letting me leave the nest too soon and fly too far, and thank you for never blaming me for forgetting to call. There are an infinite number of ways to live, thank you for letting me find my own path to become who I am today. As I grow older, I see the pieces of you in my personality, intellect, and habits. I cherish each discovery as a reminder of how present you two are in my life despite being so far away.

While being a member of a large family can be overwhelming, I have grown to appreciate the unique qualities earned through our years of rambunctious childhood. To my siblings Andrea, Lauren, Conor, Ty, and Fields, thank you for providing a dynamic and enriching environment to be raised in, and for accepting me as your brother. I cannot imagine the person I would be today without each of your influences. Thank you for spawning the next generation of our family and relieving me of the pressure to bring a child into this world—this has afforded me the opportunity to pursue this degree without feeling as though I am running out of time to start a family of my own. I will always have a great, big family back home in Mississippi, and for that I am so grateful.

To all the family who have known me forever and to all the friends I have made along the way, I hope you forgive me for my absence in your lives while completing this degree. The amount of focus devoted to this degree was due to my love for science, and a desire to make the most of the opportunities I had in front of me. With this degree completed, I hope to work less frenetically and call more regularly.

Finally, I wish to acknowledge the profound impact my soulmate, my partner, and my future husband, Mr. Phong Cam Ly, P.E., has had on my success. Your ceaseless support and undying love has healed me when I was at my lowest, has nourished me when I was too focused to eat, and has pushed me forward when I could not keep going. In all honesty, I do not think I could have completed this degree without you by my side. You wake up every morning singing, you find every excuse to laugh, you refuse to waste a moment, and you never let me second guess myself. Your presence in my life has been a gift beyond measure, and I could not be who I am today without you. There were countless ways I could have failed, but your persistent care has buoyed me over the waves and into safer harbors. I dedicate this dissertation, and all my future endeavors, to you—with all my heart and mind.

Chapter 1: Biochemistry of non-LTR R2 reverse transcriptases and OTTR

Introduction

Transposable elements are a rich source of nucleic acid enzymes with functions that can be adapted for molecular biology applications. Within the genomes of nearly every eukaryote, these elements are locked into an evolutionary arms-race with their hosts, with the active elements which remain facing continuous pressure to suppress their transcription, translation, and replication. One class of these transposable elements is the non-long terminal repeat (non-LTR) retrotransposon, a prolific group with members so successful that they account for over 20% of the human genome, *e.g.* LINE1 (Lander et al. 2001). Despite the magnitude of their representation in the human genome, only ~1 out every 500,000 copies encode the entire LINE1 gene, and of those only ~90 copies are thought to be active (Brouha et al. 2003). Characterizing and understanding the various nucleic acid processes that contribute to the successful proliferation of these elements offers fertile ground for discovering innovative biotechnological tools in molecular biology.

Of interest to this dissertation is the non-LTR R2 retrotransposon from the domesticated silkmoth, *Bombyx mori*. R2 is unique in that the entire protein complement of the ribonucleoprotein (RNP) complex is encoded in a single open reading frame (ORF), making it particularly amenable to biochemical characterization (Eickbush and Eickbush 2015). This single-ORF protein encodes multiple functions. One function combines the deoxyribonucleic acid (DNA)-binding activity from the zinc-finger and Myb domains, and an endonuclease (EN) domain, to enable the protein to recognize and nick a specific, conserved site in the 28S ribosomal (r)DNA gene (Eickbush and Eickbush 2015). A second function the ribonucleic acid (RNA)-binding ability, which enables RNA sequences encoded in the 3′ and 5′ untranslated region (UTR) to be recognized by the RNA binding motifs 0 and -1 (with respect to reverse transcriptase (RT) motifs 1 through 7) (Eickbush and Eickbush 2015; Christensen et al. 2006; Wilkinson et al. 2023). A third function of the R2 protein is the RT activity, which differs from the primer-hybridization-dependent RT activity of classically studied retroviral RTs. For R2 RTs, and other non-LTR RTs, genomic insertions are mediated through target-primed reverse transcription (TPRT), which combines the previously described DNA- and RNA-binding activities to nick a specific genomic sequence, liberating a 3′-hydroxyl (3′-OH) to prime complementary (c)DNA synthesis of the bound RNA template (Eickbush and Eickbush 2015).

In the case of R2 RT, this insertion exclusively occurs in the well-conserved 28S rDNA loci, which cells maintain at a high-copy number (Eickbush and Eickbush 2015). Remarkably, TPRT does not require any homology between the genomic nick-site and the 3′ end of the R2 RNA sequence; instead, the R2 RT extends the 3′-OH of the nicked DNA with non-templated additions (NTAs) that complement the 3′ sequence of the R2 RNA to initiate reverse transcription (Eickbush and Eickbush 2015; Deng et al. 2023).

While the disparate steps of this process remain unclear, advances in structural biology, nucleic acid biochemistry, and evolutionary analysis are helping to build an understanding of these R2 retroelements (Wilkinson et al. 2023; Deng et al. 2023; Thawani et al. 2024; Rodríguez-Vargas and Collins 2024). Despite these gaps, a recent study of bird R2 proteins has identified a new modality for genome engineering in human cells, which do not possess R2, by exploiting TPRT for *trans*-gene delivery through the co-transfection of a R2 RT messenger (m)RNA with a *trans*-gene encoding RNA flanked by the 5′ and 3′ R2 RNA sequences (Zhang et al. 2024). Indeed, further study of R2 biochemistry is warranted as these enzymes could be repurposed as tools for molecular biology research.

Biochemistry of non-LTR R2 reverse transcriptases.

First described in 1987 (Burke et al. 1987), the *B. mori* R2 RT has been the preferred prototypical R2 protein for biochemical characterization since it can be recombinantly expressed and purified as a functional protein, and is capable of being reconstituted in *in vitro* assays to study it role as a nucleic acid binding protein, an EN, and a RT (Luan et al. 1993; Luan and Eickbush 1995; Christensen and Eickbush 2004; Jamburuthugoda and Eickbush 2014). Nearly four decades later, questions still remain regarding the discrete steps necessary for R2 retroelements to complete their integration, but the following steps thought to be required are: 1) Co-transcription with rRNA by RNA Pol I from an rDNA unit; 2) liberation of the 5′ end of R2 RNA from its rRNA co-transcript by activity of the encoded 5′ ribozyme; 3) nuclear export and protein translation; 4) co-translational folding and binding to *cis*-RNA via a 3′ UTR feature; 5) nuclear import of assembled RNP complex; 6) target site honing through site-specific DNA-binding motifs; 7) DNA nicking by the EN domain; 8) cDNA reverse transcription primed from nicked DNA facilitated by template-jumping to the 3′ end of the RNA; and, yet to be clearly described, 9) second-strand synthesis and resolution of DNA lesions (Eickbush and Eickbush 2015). Compared to classically described retroviral RTs, the R2 RTs can not only initiate reverse transcription in a distinct way but also have enhanced polymerase features such as greater processivity and strand-displacement.

To date, RT encoding elements have been grouped into four supergroups depending on their N- and C-terminal domains and presence of RNase H domain: prokaryotic RTs, retrovirus-like and LTR RTs, Penelope-like and telomerase RTs, and non-LTR and RT-related cellular genes (Gladyshev and Arkhipova 2011). Retroviral RTs such as Moloney murine leukemia virus (MMLV) have been harnessed in molecular biology as tools for DNA-primer-dependent reverse transcription of RNAs with well-defined 3′ sequences, such as polyadenylated(A) mRNAs. This use is possible due to the way these retroviral genomes are replicated. Members of the retroviral clade (including MMLV, avian myeoloblastosis virus (AMV), pararetroviruses, and LTR retrotransposons) require the hybridization of a host-derived transfer (t)RNA primer to a complementarily-encoded tRNA-primer-binding site to initiate reverse transcription of the retroviral RNA genome (Marquet et al. 1995). Indeed, the presence of a tRNA primer binding site within the same RNA genome that encodes a RT with a RNase H

domain is a defining characteristic of these retroviral RTs (Marquet et al. 1995).

In others clades, the genes which encode RTs such as hepadnaviruses, telomerase, and non-LTR retroelements such as R2 do not also encode their own primer binding sites (Marquet et al. 1995). Instead, these RTs have evolved different mechanisms for initiating cDNA synthesis. Among these disparate RTs is the hepadnavirus RTs, which solved the primer initiation problem with the most extreme of means: by using a hydroxyl from an encoded tyrosine residue near its RT domain as a primer (Marquet et al. 1995). A more elaborate system can be found with telomerase, which famously uses the non-coding (nc) telomerase RNA, TER, as its own cDNA synthesis template, allowing it to extend the ends of chromosomes with the assistance of various proteins in a complex known as the telomerase holoenzyme (Jiang et al. 2013; Nguyen et al. 2018).

Non-LTR retroelements, such as R2, prime reverse transcription by EN activity, where a targeted nick in the DNA yields a 3′-OH which the R2 RT can use to prime the addition of the first cDNA nucleotide (Eickbush and Eickbush 2015). In most cases, R2 integration yields a consistent, scarless junction between the 28S rDNA and the 3′ end of the newly-inserted R2 sequence, but longer, A-rich NTAs have been observed (Luan and Eickbush 1995). Compared to retroviral RTs, like AMV RT, R2 RT is a more processive polymerase capable of transcribing longer, more full-length cDNA transcripts from longer, more-structured RNA templates (Bibiłło and Eickbush 2002b). Unlike retroviral RTs, the R2 RT is not impeded by template secondary structure as it can perform efficient strand-displacement during cDNA synthesis, facilitating the removal of annealed RNA or DNA strands from the template to maintain processivity (Bibiłło and Eickbush 2002b; Kurzynska-Kokorniak et al. 2007). While this evolved processivity comes at a cost of slowed release of the transcribed duplex from the RT protein due to high-affinity interactions (Bibiłło and Eickbush 2002b), the ability to displace RNA:RNA and RNA:DNA duplexes with unimpeded reverse transcription may explain how this RT is able to complete its unusual replication, both first and second strand synthesis, without the need of a RNase H domain (Kurzynska-Kokorniak et al. 2007). Moreover, this processivity could be useful in molecular biology applications where engineered retroviral RTs have difficulty, such as tRNA or ncRNA sequencing.

Important steps of *B. mori* R2 integration can be reconstituted in an *in vitro* TPRT reaction. In this reaction *B. mori* R2 RT is supplied with DNA containing the 28S rDNA target sequence, the R2 3′ UTR RNA sequence, deoxynucleotide triphosphates (dNTPs), salts, and buffers, which are needed for it to synthesize cDNAs composed of the DNA sequence 5′ to the nick site (*i.e.*, primer) and the reverse transcribed 3′ UTR (*i.e.*, cDNA) (Luan et al. 1993; Luan and Eickbush 1995; Christensen and Eickbush 2005). Unexpected cDNA products accumulate in these *in vitro* assays, such as cDNA molecules with a length corresponding to the reverse transcription of two or more copies of the R2 RNA (Bibiłło and Eickbush 2002a). When assayed without either the 28S rDNA target sequence or a DNA primer, the R2 RT has been reported to use the 3′-OH of one R2 3′ UTR RNA as a primer for template-dependent reverse transcription of a second R2 3′ UTR RNA, yielding a chimeric 5′-RNA-cDNA-3′ (Bibiłło and Eickbush 2002a). The accumulation of both *in vitro* side-products suggests a robust capacity for

R2 RT to transcribe cDNA, even under atypical conditions or with atypical substrates.

NTA is a shared phenomenon across RTs where extra nucleotides are added to the cDNA 3′ end once the RT has reached the 5′ end of the RNA template. NTAs persist on cDNAs as RTs have not evolved 5′ to 3′ proofreading exonuclease activity, and different RTs have varying preferences for what nucleotides are added to cDNAs as NTAs (Zhao C et al. 2018; Boutabout et al. 2001; Golinelli and Hughes 2002; Wulf et al. 2019; Chandramouly et al. 2021). In the case of *B. mori* R2 RT, the preference is for the nucleotides deoxyadenosine (A) or deoxyguanine (G), collectively deoxypurine (R) (Bibiłło and Eickbush 2002a, Upton et al. 2021), while MMLV RT prefers to use the nucleotide deoxycytidine (C) (Wulf et al. 2019). Some RTs utilize NTA at the cDNA 3′ end to facilitate continuous synthesis from discontinuous templates, employing the NTA as a primer for “template jumping.” During *in vivo* MMLV replication, this activity is vital and additionally requires the MMLV RNase H domain to condition the nascent cDNA for template jumping by degrading the 5′ end of the reverse transcribed template RNA so that the cDNA may prime reverse transcription of a second viral RNA template (Garcés and Wittek 1991). In molecular biology, MMLV has been used to synthesize a continuous cDNA from two separate templates in a process reliant on a +3 CCC NTA to the cDNA of the first template to support a template jump to a second 3′-rGrGrG template (Wulf et al. 2019).

Similar template-jumping behavior can be used by some RTs to initiate cDNA synthesis without the need for DNA primer hybridization, instead relying on a single base-pair between primer and template. Both thermostable group II intron RT and *B. mori* R2 RT can use a DNA/RNA primer-duplex with a single 3′-overhanging nucleotide (hereafter +1 N, *i.e.,* N is A, C, G or deoxythymidine (T)) to initiate cDNA synthesis on any RNA template 3′ end (Xu et al. 2019, Upton et al. 2021, Pimentel et al 2022). For *B. mori* R2 RT, robust template-jumping activity is routinely observed where NTA to the cDNA forms a new +1 primer to facilitate synthesis of concatemeric cDNA using multiple templates (Upton et al. 2021; Pimentel et al. 2022). This simplicity enables robust, seamless copying of multiple RNAs end-to-end by R2 RT (Upton et al. 2021).

In studies of other DNA polymerase, it has been found that *in vivo* polymerase assay using manganese instead of magnesium can relax the specificity of polymerases, unmasking useful and new activities such as terminal transferase (Vashishtha et al. 2016). Manganese-dependent terminal-transferase activity has been described for DNA polymerases Polλ, Polμ, PolΘ, with crystallographic evidence for manganese-selective terminal transferase activity reported in Polβ (reviewed in Balint and Unk 2023; Kent et al. 2016; Pelletier et al. 1996). Recently, R2 RT has been found to have toggleable terminal-transferase activity in the presence of manganese, capable of using either dNTP or dideoxynucleotide triphosphates (ddNTP) with a strong R bias (Upton et al. 2021). Similarly, recent work in group II intron-like RT *Pseudomonas aeruginosa* and group II intron RT *Eubacterium rectale* have demonstrated both RNA and ssDNA terminal transferase activity in the presence of manganese (Upton et al. 2021, Park et al. 2022) with differences in nucleotide and substrate preferences and tailing lengths. For R2 RT, manganese-dependent terminal-transferase activity has been observed for single-stranded and double-stranded RNA and DNA, as well as for DNA/RNA duplexes

(Upton et al. 2021). Moreover, *B. mori* R2 RT can robustly terminally transfer a ddNTP to a nucleic acid 3′ end, with a preference for dideoxyadenosine triphosphate (ddATP) or dideoxyguanine triphosphate (ddGTP) (Upton et al. 2021), a potentially useful molecular biology approach to standardize the 3′ end sequence of a pool of nucleic acids to a common, single nucleotide.

The invention of OTTR.

In 1977, fellow Johnian (*i.e.,* member of St. John's College, Cambridge) and twice recipient of the Nobel Prize in Chemistry, Fredrick Sanger, established the first DNA sequencing method by dideoxy chain-termination, enabling the human mitochondria and bacteriophage λ genomes to be sequenced. From this, the human genome project was possible, ushering in the genomics era, a time in which this dissertation was completed.

Generally, RNA sequencing requires the primer-hybridization-dependent activity of retroviral RTs to convert RNA to cDNA. A challenge for RNA sequencing is the overwhelming abundance of rRNA, which is ~90% of the total RNA population. Poly(A) messenger RNA (mRNA) sequencing overcomes this burden by enriching for mRNAs through the selective hybridization of poly(T) DNA primer oligonucleotides to initiate cDNA synthesis by retroviral RTs. While reverse transcription is still required, modern, high-throughput sequencing methods that quantify gene expression by sequencing mRNA transcripts (hereafter mRNA sequencing) need to synthesize cDNAs with additional 5′ and 3′ "sequencing" adapters before the cDNAs can be amplified and sequenced. Depending on the application, a common approach for mRNA sequencing is to enrich for poly(A) mRNA transcripts, fragment into ~200 nucleotide pieces, add 3′ poly(A) tails to each fragment with poly(A) polymerase, and reverse transcribe with retroviral RTs from a hybridized poly(T) DNA primer, which includes the 5′ sequencing adapter. Following reverse transcription, the 3′ sequencing adapter can either be ligated to the cDNA, or in some cases, a RT like MMLV can extend the cDNA with multiple C NTAs to facilitate a template jump to an adapter template, as described above (Wulf et al. 2019). These "cDNA libraries" are amplified by DNA primers that anneal to the introduced 5′ and 3′ sequencing adapters (*e.g.,* Illumina TruSeq primers Read1 and Read2) and extend the final library with identifying "index" sequences (*e.g.,* Illumina TruSeq indices i5 and i7) and flow-cell annealing sequences (*e.g.,* Illumina P5 and P7). While over a billion sequencing reads could easily be recovered from a single mRNA sequencing library, it is typical to only need 10 – 100 million reads per replicate as of the writing of this dissertation.

For longer mRNAs these approaches work well, enabling experiments as sensitive as single-cell mRNA sequencing (Vandereyken et al. 2023); however, this success has not perfectly translated to small and ncRNA (Tosar et al. 2024). Small RNAs like micro(mi)RNAs, tRNAs or tRNA-derived fragments, glycosylated Y-RNAs, and other ncRNAs are important cellular signals and post-transcriptional and epigenetic regulators of the cell, worthy of being quantified in parallel with mRNA sequencing

experiments (Muthukumar et al. 2024, Shi et al. 2022, O'Brien et al. 2020, Flynn et al. 2021, Tosar et al. 2024). One challenge to small RNA quantification is that the 3′ sequences are inconsistent (*i.e.,* no poly(A) tail), making DNA primer hybridization and retroviral reverse transcription impossible without first relying on RNA ligases to covalently attach a sequencing adapter to the RNA 3′ end. Unfortunately, these ligases jeopardize the reliability of small RNA quantification by introducing sequencing bias through the preferential ligation of certain RNAs to the adapter (Fuchs et al. 2015, Tunney et al. 2018).

In mRNA sequencing, ligase bias is not generally considered a major problem in library preparation; after all, many discrete RNA fragments can come from a single mRNA transcript, and the calculated expression estimates are informed by the coverage across the entire gene body. For small RNAs, the entire transcript may only give rise to a single, full-length read, which may additionally be under- or over-represented in the final sequencing library simply due to library preparation ligase preferences (Fuchs et al. 2015, Tunney et al. 2018, Ferguson et al. 2023). Beyond this bias, these small RNA, or RNAs digested by nucleases from cross-linked immunoprecipitation (CLiP) experiments or formalin-fixed paraffin-embedded (FFPE) tissues, may require 3′ end repair by T4 polynucleotide kinase or other phosphatases (Tosar et al. 2024). Inconsistencies in enzymatic activity for the 3′ end repairs steps or ligation can jeopardize small RNA library preparation before reverse transcription begins. While approaches such as poly(A) polymerases can be used in place of ligases for small RNAs, the same issues with substrate preferences remain (Yehudai-Resheff and Schuster 2000). Beyond ligation, tRNAs, other ncRNAs, and nucleic acids from CLiP experiments or FFPE tissues can have post-transcriptional nucleotide modifications which retroviral RTs cannot reverse transcribe, pre-maturely stopping cDNA without completely representing the transcript sequence end-to-end (Gustafsson et al. 2022, Flynn et al. 2021). To address these issues, advances in molecular biology would be needed.

Recently, a ligation-free small RNA library preparation method, Ordered Two-Template Relay (OTTR), was invented by the Dr. Heather E. Upton, Ms. Sydney Pimentel, Prof. Kathleen Collins, and the author of this dissertation. This method took advantage of the previously described activities of the *B. mori* R2 RT to enable the capture of small RNA fragments in a single-tube, 4-hour workflow (Upton et al. 2021). This method relied on BoMoC(ed) (or shorthand BoMoC), a recombinantly expressed *B. mori* R2 RT lacking the N-terminal linker, zinc-finger, Myb DNA binding domains, and with an N-terminal MBP and C-terminal 6X-histidine tags and an inactivated EN domain (D996A) (Upton et al. 2021). BoMoC can be expressed and purified at a high concentration following a 3-step nickel affinity, heparin, and size-exclusion chromatography steps, and is highly active in *in vitro* assays (Upton et al. 2021).

The invention of OTTR required careful evaluation of BoMoC substrate preference for two specific reactions: terminal transferase (hereafter 3′ labeling) and template jumping. Based on 3′-labeling assays with various substrates carried out with BoMoC and EuRe, a recombinant *E. rectale* group II intron RT with N-terminal MBP and C-terminal 6X-histidine tags, it was clear BoMoC had exceptional 3′-labeling competency, capable of labeling ssRNA, ssDNA, double-stranded (ds)RNA, dsDNA,

and RNA/DNA duplexes in the presence of manganese (Upton et al. 2021). While EuRe also demonstrated manganese-stimulated 3′-labeling, which had not been previously reported for a group II intron RTs at the time of invention (Park et al. 2022), it had a preference to 3′ label ssDNA and dsDNA with some, but an overall dramatically reduced, activity for ssRNA, dsRNA, and RNA/DNA duplexes (Upton et al. 2021). The 3′-labeling activity of BoMoC was notably higher with dATP compared to other dNTPs, with a slightly reduced temperature of 30 °C and longer incubation capable of improving the labeling outcome. Advantageously, BoMoC was also capable of using ddATP to complete 3′-tailing, enabling the addition of single nucleotides to the 3′ ends of ssRNA, dsRNA, ssRNA, or dsDNAs (Upton et al 2021).

BoMoC was found to have preferences for initiating cDNA synthesis, and subsequent template jumps (Upton et al 2021). Using blunted or short +1 or +2 DNA/RNA primer duplexes, BoMoC could be stimulated to initiate cDNA synthesis with an RNA template, but with longer duplexes this was disfavored (Upton et al. 2021, Pimentel et al. 2022). For this reason, template jumping was sensitive to conditions that modulated the number of cDNA NTAs. Under certain dNTP mixes, excessive NTAs would accumulate at the ends of cDNA duplexes such that subsequent template jumping would be impossible by BoMoC (Upton et al. 2021). In other conditions, NTAs can be attenuated such that dozens of template jumps could accumulate, yielding a characteristic “laddering” profile of concatemeric cDNAs when resolved by denaturing (d) urea polyacrylamide gel electrophoresis (PAGE) (Upton et al. 2021).

From these two enzymatic activities, a new strategy for small RNA library preparation was established (Upton et al. 2021). Since BoMoC preferred to 3′ label RNAs with ddATP (and ddGTP) in the presence of manganese, it would be possible to use this activity to standardized the 3′ ends of the input RNAs with 3′ ddR. Based on BoMoC preferences, +1 deoxypyrimidine (Y, *i.e.*, T or C) primer duplexes that encoded the Illumina TruSeq Read2 sequencing adapter were chosen to capture these 3′ ddR labeled RNAs and prime cDNA synthesis in the presence of magnesium. From these cDNA duplexes, a +1 G NTA would be preferred, forming a new primer sufficient to stimulate a second template jump to a 3′C adapter template that encoded the Read1 TruSeq sequencing adapter. To discourage further jumping, large and fluorescent polymerase blocks were introduced on the 5′ end of the adapter template, as well as the 5′ end of the primer duplex. From two discontinuous templates, a continuous cDNA would be synthesized from this single-tube workflow in four hours instead of the days of effort and multiple experiments traditionally required for small RNA library preparation (Upton et al. 2021).

The name for OTTR was chosen to describe both the order of jumping between these templates and the number of jumps taken by BoMoC (Upton et al. 2021). As originally described for miRNAs, this method was found to have the lowest levels of detectable sequencing bias when comparing libraries prepared by other commercial small RNA sequencing methods using a common reference standard, the miRXplore miRNA equimolar pool of 962 miRNAs (Upton et al. 2021). This was surprising as the reliance on 3′ labeling used in OTTR might have yielded similar issues faced by library preparation kits like SMARTer-seq small RNA library kit, which used a poly(A)

polymerase to standardized RNA 3′ ends with longer A-tails (Coenen-Stass et al. 2018). Unlike SMARTer-seq, however, OTTR relied on template jumping by the formation of a single nucleotide base-pair between the primer duplex and 3′ddR-labeled template, not poly-T hybridization, to initiate reverse transcription (Upton et al. 2021, Coenen-Stass et al. 2018). In the case of OTTR, it may be possible for templates that did not receive a 3′ ddR label to be captured, which would not be possible for RNAs left without a poly(A) tail in SMARTer-seq (Upton et al. 2021).

When directly compared by a random forest regression models trained on 5′ and 3′ sequence, other methods like TGIRT-seq, which used a similar but distinct +1 N overhang primer duplex and a thermostable group II intron to capture the 3′ ends of RNA templates, a detectable 3′ G/C bias was introduced, as well as a 5′ ligase-dependent bias signature (Upton et al. 2021, Xu et al. 2019). OTTR, on the other hand, did not suffer from the same biases, instead yielding relatively uniform capture for both 5′ and 3′ ends despite using a +1 Y primer duplex. Overall, the OTTR method appeared to deliver a low-bias small RNA profile and was capable of transcribing libraries from low-input samples (Upton et al. 2021), but had yet to be optimized or extensively applied. In subsequent chapters, this dissertation will detail optimizations which improved OTTR for small RNA library preparation, the development of a simplified and improved ribosome profiling method which took full advantage of OTTR, and lastly, future improvements and applications of this method.

Chapter 2: Improved precision, sensitivity, and adaptability of Ordered Two-Template Relay cDNA library preparation for RNA sequencing

Abstract

Sequencing of RNAs that are biologically processed and/or degraded to less than ~100 nucleotides typically involves the use of multi-step, low-yield protocols with bias or information loss inherent to ligation and polynucleotide tailing steps. We recently introduced Ordered Two-Template Relay (OTTR), a small RNA sequencing method that appends 5′ and 3′ sequencing adapters to the end-to-end capture of input-molecule sequence during reverse transcription. Previous studies establish that OTTR produces microRNA, tRNA or tRNA-derived fragments, and ribosome-protected messenger RNA footprint libraries with low bias. Here we sought to characterize, quantify, and ameliorate remaining sources of bias and imprecision of OTTR in the capture of end-to-end RNA sequences. We introduce new metrics for the evaluation of sequence capture and use them to optimize buffers, reverse transcriptase sequence, adapter oligonucleotide design, and overall workflow to improve the OTTR method. Modifications of the reverse transcriptase and adapter oligonucleotides increased the 3′ and 5′ end-precision of sequence capture and minimized overall library bias. Improvements in the protein purification process reduced non-productive sequencing reads by minimizing bacterial nucleic acids that compete with low-input RNA molecules for cDNA synthesis, bringing the required amount of input sample that gives less than 10% bacterial nucleic acid sequencing reads down to roughly 3 picograms (less than 1 fmol) of microRNA. Overall, this work informs considerations for the unbiased end-to-end capture and annotation of RNAs independent of their sequence, structure, or post-transcriptional modifications.

Introduction

Nearly two decades of innovations in polyadenylated messenger (m) ribonucleic acid (RNA) sequencing have generated numerous methods for the bulk and single-cell sequencing of protein-coding transcripts (Vandereyken et al. 2023). In comparison, methods that capture the full complexity of non-coding RNAs (ncRNA) or processed small RNAs remain underdeveloped (Tosar et al. 2024) and comprehensively profiling these RNAs at low input, for example single-cell profiling, is an unmet challenge (Wang et al. 2019, Hücker et al. 2021). Although mRNA sequencing is an unquestionably useful tool for assessing cell state, small RNA profiles of microRNAs (miRNA), transfer RNAs (tRNA), tRNA-derived fragments (tRF), vesicle-trafficked and lysosomal RNA fragments, and non-polyadenylated pathogen-derived RNAs could be additionally or even more informative (Muthukumar et al. 2024, Shi et al. 2022, O'Brien et al. 2020). Increasingly, ncRNAs are recognized as important pre- and post-transcriptional regulators of gene expression, with emerging roles in epigenetic inheritance and cell-to-cell communication (Kaikkonen et al. 2011, Boskovic et al. 2020, van Niel et al. 2022).

Without widespread adoption of sensitive, accurate, and reproducible methods to inventory these transcripts, the quantitative appraisal of biology is incomplete.

To fill this knowledge gap, complementary (c) deoxyribonucleic acid (DNA) library preparation methods that do not require specific RNA sequence motifs and are not impaired by RNA structure or modified ribonucleotides are needed. Retroviral reverse transcriptases (RTs) exclusively initiate cDNA synthesis by annealing a cognate primer to a sequence in the RNA template, *e.g.,* in polydeoxythymidine primed cDNA synthesis from mRNA. To make cDNA libraries from small RNAs, which do not possess a shared 3′ sequence, the 3′ ends of RNAs are first either tailed with adenosines by enzymes such as *Escherichia coli* or yeast polyadenosine polymerase (yPAP), or ligated to a RNA adapter using specialized ligase enzymes (Lu et al. 2007, Coenen-Stass et al. 2018, Androvic et al 2022). Both 3′-standardization methods can be inefficient and favor the capture of specific RNAs based on sequence or structure (Yehudai-Resheff and Schuster 2000, Fuchs et al. 2015, Tunney et al. 2018, Ferguson et al. 2023). Another approach to extend a cDNA with an adapter sequence is to rely on RT activity to extend of the 3′ ends of cDNAs with multiple non-templated nucleotide additions (NTA) to enable base-pairing with an oligonucleotide adapter template (hereafter adapter template). As first developed for cDNA library synthesis using the retroviral Moloney murine leukemia virus RT, these NTA nucleotides support base pairing with an adapter template used to enable continued cDNA synthesis to incorporate an adapter sequence (Wulf et al. 2019, Zhu et al. 2001).

A similar but distinct “template jumping” activity has been observed using RTs outside the retroviral clade, which can initiate cDNA synthesis from a blunt-ended primer duplex or one with a single-nucleotide 3′ overhang. This is exploited in thermostable group II intron RT cDNA library synthesis, which uses a DNA/RNA primer duplex with a single, 3′-overhanging, degenerate deoxyribonucleoside (N, *i.e.*, deoxyadenosine (A), deoxythymidine (T), deoxyguanosine (G), or deoxycytidine (C)) to capture RNAs without a standardized 3′ sequence (Xu et al. 2019). Recently, we exploited the particularly robust template-jumping activity of the non-long terminal repeat (non-LTR) retrotransposon protein encoded by *Bombyx mori* R2 (Bibiłło and Eickbush 2002a, Bibiłło and Eickbush 2002b) to establish a new cDNA library synthesis method (Upton et al. 2021). The *B. mori* R2 RT can initiate reverse transcription on RNA or DNA templates physically separate from a primer duplex with the primer 3′ end blunt or with a +1 or +2 3′-overhang (hereafter +1 or +2), and synthesize concatemeric cDNA by completing multiple, back-to-back template jumps (Bibiłło and Eickbush 2002a, Upton et al. 2021, Pimentel et al. 2022). The capacity to initiate reverse transcription in this way may have evolved with the non-LTR retrotransposon insertion mechanism of target-primed reverse transcription—in this process the retrotransposon-protein endonuclease (EN) domain nicks the target site, liberating a 3′-hydroxyl (3′-OH) primer that the RT domain uses to begin cDNA polymerization from the 3′ end of its bound RNA (Eickbush and Eickbush 2015). Non-LTR retrotransposon RTs have also evolved efficient strand-displacement activity to confer the high processivity needed to successfully copy a long and structured retrotransposon RNA template (Bibiłło and Eickbush 2002b, Kurzynska-Kokorniak et al. 2007).

Using the *B. mori* R2 protein, we established Ordered Two-Template Relay (OTTR) to reverse transcribe a continuous cDNA from discontinuous templates, notably using two templates in a specific order that ensures each input-template cDNA is flanked on both sides by the Illumina TruSeq sequencing adapters, Read1 and Read2 (Upton et al. 2021). OTTR exploits the manganese-stimulated terminal-transferase (hereafter 3′-labeling) activity of a truncated, EN-inactivated *B. mori* R2 protein (hereafter BoMoC), using a dideoxynucleotide triphosphate (ddNTP) to extend the 3′ ends of single-stranded (ss) or double-stranded DNAs, RNAs, or DNA/RNA duplexes (Upton et al. 2021). While other polymerases have also demonstrated manganese-enhanced 3′-labeling activity (Pelletier et al. 1996, Kent et al. 2016, Park et al. 2022, Ohtsubo et al. 2017, Balint and Unk 2024), the divalent-ion-dependent toggle of R2 protein activity from exclusively 3′ labeling in manganese to exclusively templated cDNA synthesis in magnesium is particularly enabling (Upton et al. 2021).

In the first step of OTTR, BoMoC is used in 3′-labeling conditions to append a single nucleotide, either dideoxyadenosine (ddA) or dideoxygaunine (ddG), collectively dideoxypurine (ddR, *i.e.*, R is either A or G), to the 3′ ends of input nucleic acids (**Fig. 2.1**, step 1). Unincorporated ddNTPs are quenched by recombinant shrimp alkaline phosphatase (rSAP), and manganese is chelated by the addition of glycol-bis(β-aminoethyl ether)-N,N,N′,N′-tetraacetic acid (EGTA) to protect against zinc-mediated RNA hydrolysis during rSAP heat-inactivation (**Fig. 2.1**, step 2). Then magnesium is added and BoMoC-mediated cDNA reverse transcription is primed by the formation of a single base pair between the 3′ ends of +1 deoxypyrimidine (Y, *i.e.,* Y is either T or C) DNA/RNA primer duplexes and the 3′ ends of 3′ddR-labeled templates (**Fig. 2.1**, step 3 1st jump). Following cDNA synthesis across the 3′ddR-labeled templates, BoMoC will extend the cDNA/RNA duplex with a +1 G NTA, providing specificity for the next template jump to the 3′ cytidine (rC) adapter template (**Fig. 2.1**, step 3 2nd jump). These template jumps convert the input pool into end-to-end transcribed cDNA with flanking 5′ and 3′ adapters.

To date, we have established OTTR as a low-bias cDNA library synthesis method for the high-throughput profiling of miRNA, extracellular RNA, and ribosome-protected mRNA footprint (RPF) sequences (Upton et al. 2021, Ferguson et al. 2023), while others have demonstrated the utility of OTTR for detecting and quantifying the relative abundance of tRNAs and tRFs and their post-transcriptional modifications (Gustafsson et al. 2022, d'Almeida et al. 2023, Manning et al. 2024). However, we have yet to establish metrics to assess precision of end-to-end sequence capture (*i.e.,* how frequently sequence capture truncates or adds a terminal nucleotide(s)), determine enzyme and buffer stability with storage, query the lower-input limit for productive library synthesis, include a unique molecular identifier (UMI) used to uniquely tag each cDNA molecule, or explore improvements of BoMoC activities via mutagenesis. In this study, we sought to identify sources of bias and information loss in the OTTR protocol and introduce solutions to mitigate these and other concerns.

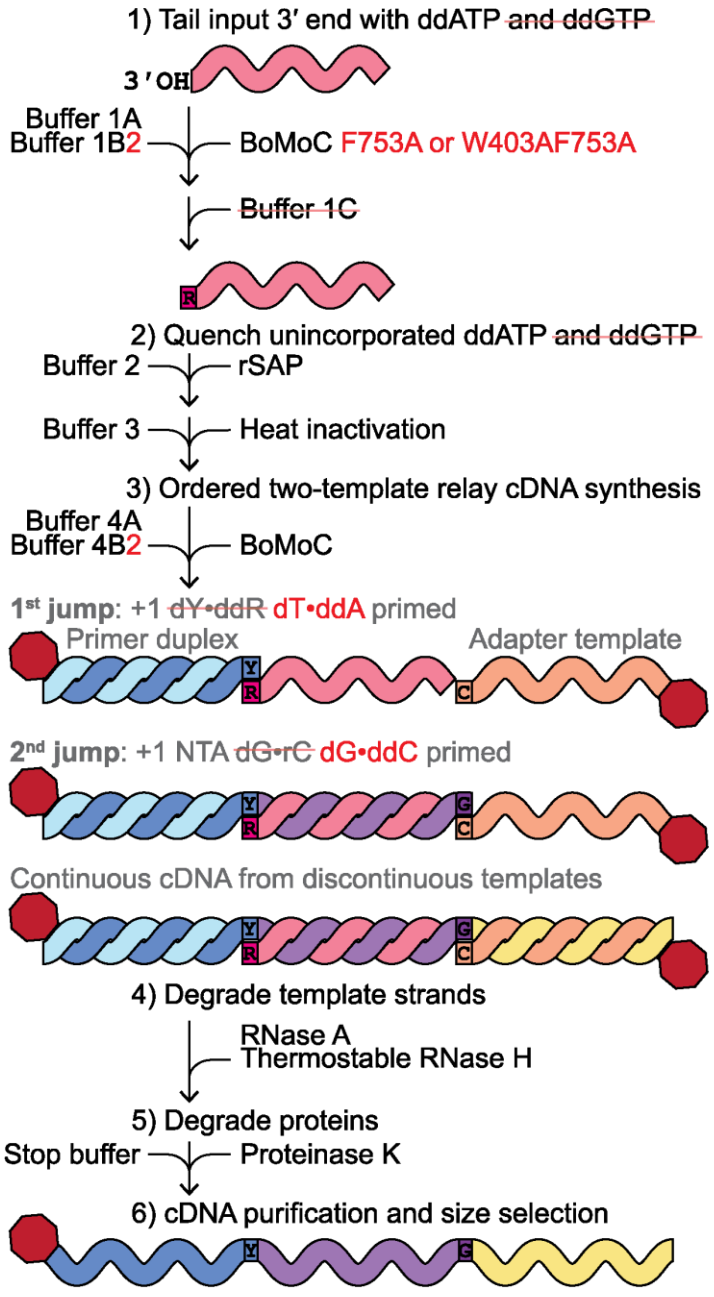


Figure legend

- Unlabeled template
- ddA or ddG labeled template
- 5′ RT jump block (Cy5 or IRD800)
- +1 Y +1 T primer DNA (Read2)
- Primer complement RNA (cRead2)
- Input template cDNA
- Template cDNA w/3′G NTA
- 3′rC 3′ddC adapter template (Read1)
- Adapter template cDNA (cRead1)

Buffer recipes

Input (9 10 µL)
0.9 pmole template in water

Buffer 1A (2 µL)
140 mM Tris-HCl pH 7.5, 1 M KCl,
14 mM DTT, 35% PEG-90000

Buffer 1B2 (1 µL)
3.5 mM ddATP, 28 mM $MnCl_2$,
28 mM $(NH_4)_2SO_4$, 10 mM sodium
acetate pH 5.5

Buffer 1C (1 µL)
3.5 mM ddGTP

Buffer 2 (1 µL)
80 mM $MgCl_2$

Buffer 3 (1 µL)
100 mM EGTA

Buffer 4A (1 µL)
10 mM $MgCl_2$, 900 mM KCl,
40% PEG-6000, 4 mM dGTP,
3 mM dDAP-TP, 0.8 mM dCTP,
0.8 mM dTTP, 0.04 mM dATP

Buffer 4B2 (1 µL)
1.8 µM +1 Y +1 T DNA/RNA
primer duplex, 3.6 µM
3′ddCRNNNNN adapter template

Stop buffer (30 µL)
50 mM Tris HCl pH 8.0,
20 mM EDTA, 0.1% SDS

Enzyme working stock (1 µL steps 1 & 3)
10 µM (1.6 mg/ml) BoMoC,
20 mM Tris-HCl pH 7.5,
320 mM KCl, 320 mM $(NH_4)_2SO_4$,
0.16 mM TCEP, 1 mM DTT,
5 mM HEPES-KOH pH 8.0,
42% glycerol

Figure 2.1: OTTR library synthesis

Original and improved OTTR cDNA library synthesis workflows. The left side explains the steps, while the right side gives buffer recipes. Modifications are highlighted in red. A stop sign indicates a modification to prevent template jumping after cDNA synthesis to the oligonucleotide 5′ end.

Results and Discussion

Precision of sequence capture at input RNA 3′ ends

In OTTR, an input RNA 3′-OH is first extended by BoMoC to have a 3′ddR nucleotide using either a dideoxyadenosine triphosphate (ddATP) or dideoxyguanosine triphosphate (ddGTP) in the 3′-labeling step before a primer duplex with cognate, single-nucleotide 3′Y overhang initiates reverse transcription (**Fig. 2.1**). To quantify the extent to which the 3′-labeling step determines input RNA conversion to cDNA, we performed OTTR either without the 3′-labeling step or using yPAP and ddATP for 3′ labeling (Martin and Keller 1998) instead of BoMoC. To assess library bias in OTTR, we made cDNA libraries from the miRXplore equimolar pool of 962 synthetic miRNAs, a reference standard commonly used to benchmark small RNA library synthesis protocols (Coenen-Stass et al. 2018, Shore et al. 2019, Xu et al. 2019, Giraldez et al. 2018, Herbert et al. 2020, Upton et al. 2021). Deviation from equimolar miRNA representation in the sequencing reads was quantified by the library-wide coefficient of variation (CV), a measure of the variation in number of sequencing reads per miRNA from the expected equimolar representation. As a metric for 3′ precision, we scored the fraction of reads for each miRNA that reflected 3′ priming on the added 3′ddR nucleotide instead of an alternative mechanism (**Fig. 2.2A**). The significance of the ddR 3′-labeling step was evident in the dramatic reduction of both library CV and 3′ precision (mean fraction of 3′-precise alignments across all miRNAs) in the cDNA libraries made without the ddR 3′-labeling step (“Unlabeled"), or when yPAP was used in place of BoMoC (**Fig. 2.2B**). This bias was most pronounced when comparing miRNAs with a 3′-adenosine (rA) or 3′-guanosine (rG) ribonucleotide (rR, *i.e.*, either rA or rG), which were over-represented, to miRNAs with a 3′-uridine (rU) or 3′-cytidine (rC) ribonucleotide (rY, *i.e.*, either rU or rC), which were under-represented (**Extended Data 2.1A-B**).

Given the poor prognosis on library capture bias when input 3′ labeling was omitted, we were surprised that all but one of the miRNA sequences were captured, with only 27 miRNAs under-represented 100-fold or more from equimolar capture. To investigate how miRNAs were captured into an OTTR cDNA library without 3′ labeling, we compared rank-order of miRNA read abundance relative to 3′ precision. Three types of non-canonical 3′-end capture were detected that resulted in a miRNA 3′ truncation, and we inferred a fourth type indistinguishable from precise 3′-end capture (**Fig. 2.2A**). First, miRNAs with a 3′rR do not need 3′ labeling to be captured. Input miRNAs with 3′rR that failed to gain a 3′ddR prior to cDNA synthesis have their 3′ ribonucleotide removed by sequence trimming and appear 1 nucleotide shorter in length (**Fig. 2.2A, ii**). In

addition, these 3′rR miRNAs become over-represented when 3′ labeling was excluded, with a corresponding under-representation of 3′rY miRNA (**Extended Data 2.1A-B**). For example, when 3′ labeling was excluded, mmu-miR-33-5p changed in rank order from 763rd to 117th, with the trimmed reads appearing as if the miRNA ended at its penultimate rC rather than 3′rA (**Fig. 2.2C**). We designated these events as imprecision from “over-capture” (**Fig. 2.2A**, imprecision mechanism **ii**).

Additional classes of non-canonical 3′-end capture were inferred from analysis of miRNAs with 3′rY that showed apparently precise capture despite omission of a 3′-labeling step (**Extended Data 2.1B**). In these cases, 3′rY capture was facilitated due to additional NTA to the +1 Y primer duplex, most likely a purine deoxynucleotide (R) NTA based on previous studies of BoMoC (Upton et al. 2021). NTA to the primer duplex thus forms a +2 YR (**Fig. 2.2A**, **iii** and **iv**) that BoMoC can use (Upton et al. 2021, Pimentel et al. 2022) to initiate a template jump. For example, miRNAs hsa-miR-599 (**Fig. 2.2D**) and mghv-miR-M1-3-3p (**Fig. 2.2E**) appeared to be precisely captured when 3′ labeling was excluded. In addition to the apparently precise capture by use of +2 YR primer (**Fig. 2.2A,** imprecision mechanism **iii**), for some miRNAs there was imprecise capture from +2 YR primer corresponding to cDNA synthesis initiation on the penultimate ribonucleotide (-1 from the 3′ end, see **Fig. 2.2E**; product by **Fig. 2.2A** imprecision mechanism **iv**). The penultimate rA also appeared to be captured, albeit inefficiently, by primer duplexes with the +1 T (**Fig. 2.2A,** imprecision mechanism **v**, resulting in an apparent two-nucleotide truncation (**Fig. 2.2D**). Previous template-jumping assays with full-length *B. mori* R2 protein detected cDNA synthesis initiation internal to a template 3′ end when precise 3′-end capture was disfavored by skew of in the deoxynucleotide triphosphates (dNTP) concentrations (Bibiłło and Eickbush 2004).

We concluded that BoMoC will capture input RNA sequences with imprecision if the 3′ddR is not present, resulting in detection of more RNA sequences than we anticipated for use of the +1 Y primer duplex (**Extended Data 2.1A-B**). While RNAs with 3′rR can be captured efficiently without a 3′ddR label, their relative abundances compared to 3′rY RNAs are inflated if the 3′-labeling step was inefficient, creating a 3′-end-dependent bias of template capture. Moreover, the misappropriation of the +1 Y primer duplex to capture unlabeled RNAs increased the number of alignments with an apparent one-nucleotide, or greater, 3′ truncation due to adapter trimming, *i.e.,* a 3′ imprecision. While this 3′ bias and imprecision can be detected if known sequences are used as input templates, we reasoned that optimization of the 3′-labeling step of OTTR would reduce sequence bias and improve the interpretation of end-to-end sequence of input RNAs that lack a reference standard.

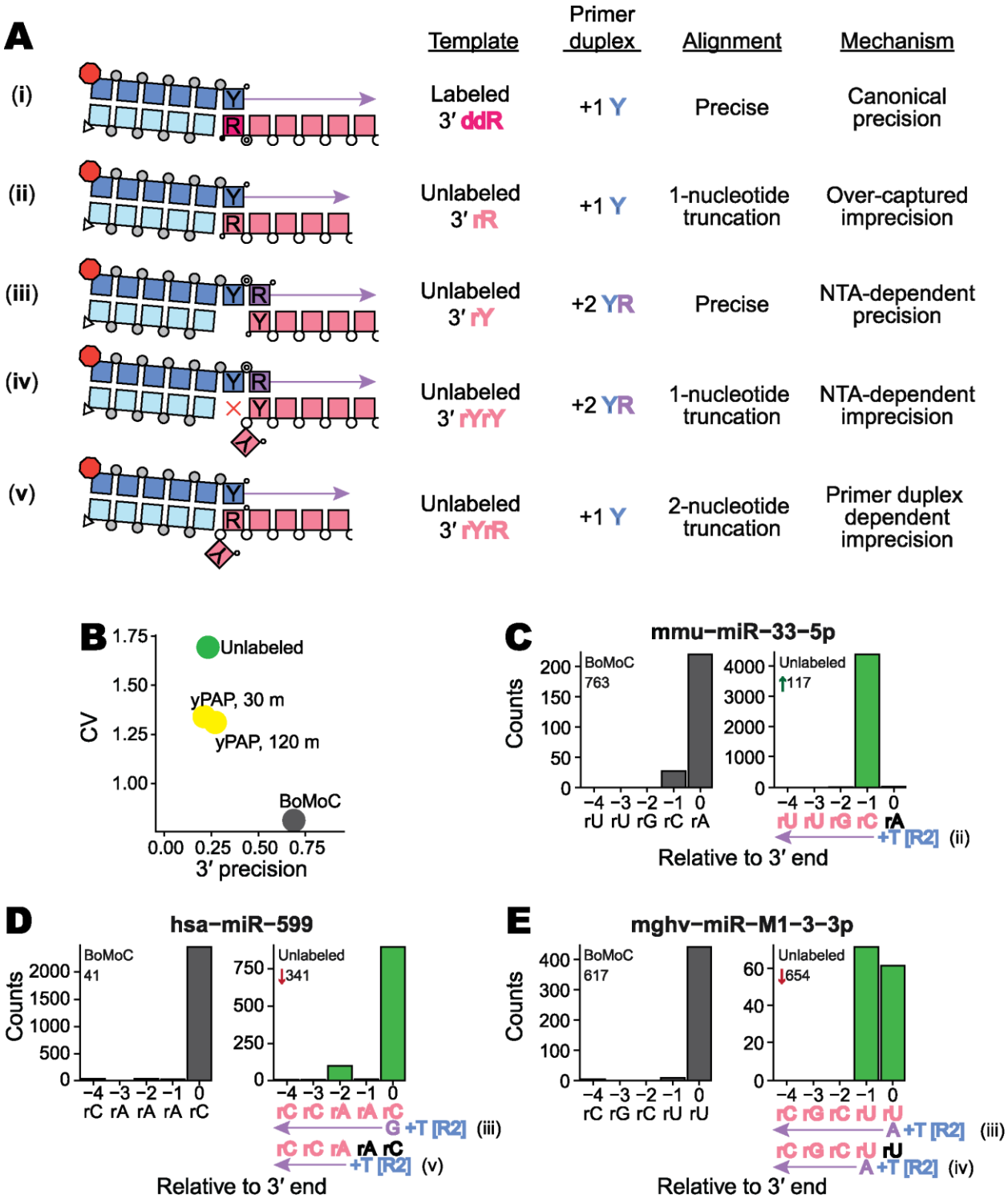


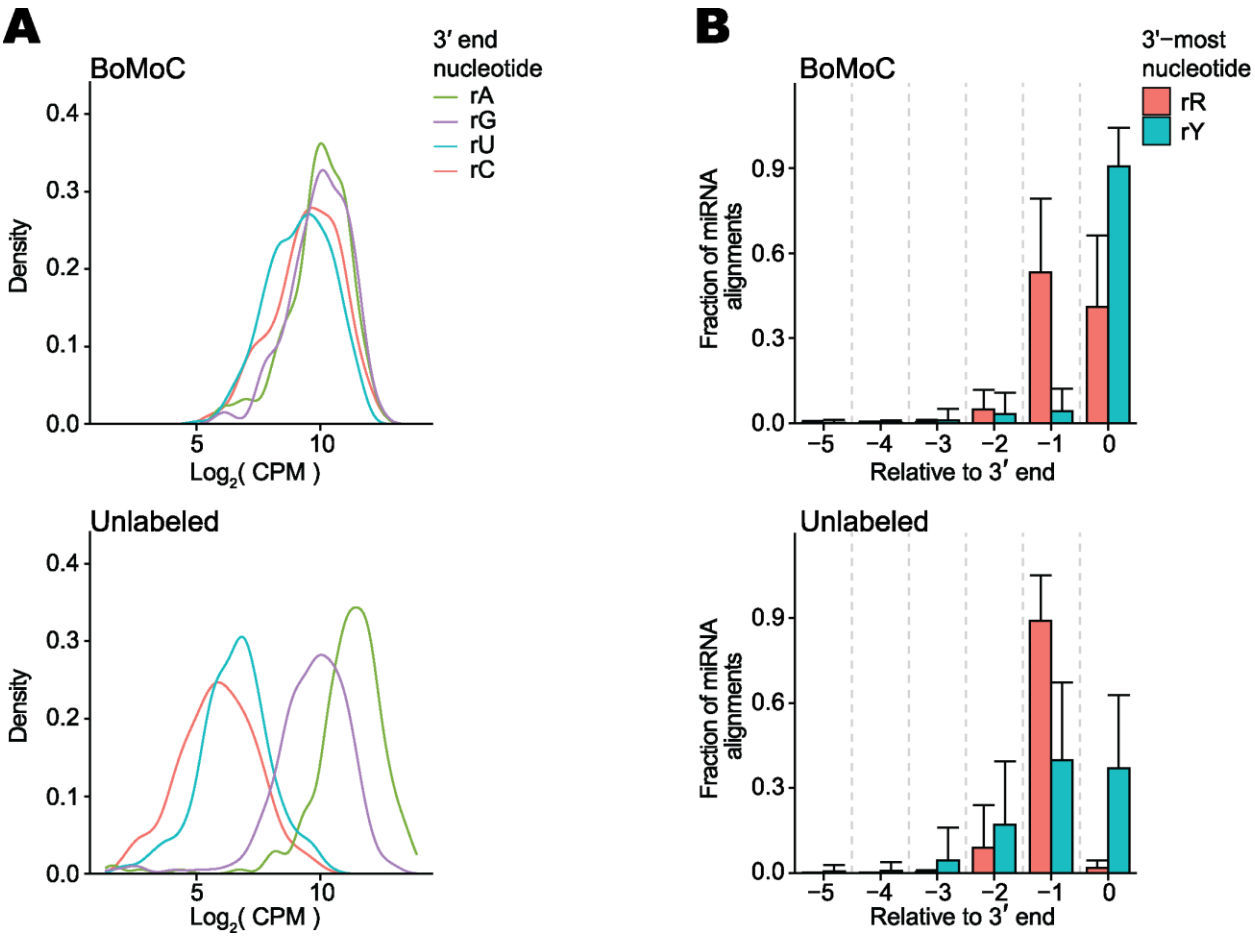
Figure 2.2: Evaluating imprecision of end-to-end sequence capture at RNA 3′-ends.

A, Mechanisms, numbered by Roman numeral, of 3′-end capture in OTTR. Column one illustrates five methods for capturing an input-template 3′ end, following the color legend

from **Figure 2.1**. Phosphodiester bonds are represented by circles: gray for chemically synthesized bonds, concentric double circle for BoMoC polymerized, and unfilled for input templates. Column two specifies the labeling status of the input-template 3′ end. Column three identifies the primer-duplex 3′ overhang supporting a template jump. Column four details the impact on 3′ sequence coverage. Column five categorizes the 3′-end capture mechanism.

B, Library-wide CV of miRXplore miRNA counts and mean fraction of alignments with complete 3′-end coverage (3′ precision). BoMoC, yPAP (for 30 or 120 minutes), or no enzyme (Unlabeled) was used for 3′ labeling.

C-E, Panel pairs depict 3′-end coverage with BoMoC (left) or no enzyme (right) for 3′ labeling. The miRNA sequence and name, and relative rank-order, are indicated at top. At bottom, position “0” is the miRNA 3′ end. Right panels have at bottom the miRNA sequence color-coded by alignment inclusion (pink) or exclusion (black). The Read2 (R2) primer sequence and adjacent +1 T overhang (+T) are in blue, and reference to the capture mechanism detailed in (**A**) is indicated. Input RNA 3′ sequence excluded by adapter trimming is indicated in black. The miRNA mmu-miR-33-5p is miRBase: MIMAT0000667, the miRNA hsa-miR-599 is miRBase: MIMAT0003267, and the miRNA mghv-miR-M1-3-3p is miRBase: MIMAT0001566.



Extended Data 2.1: Comparison of 3′ precision with and without the step of input RNA 3′ labeling.

A, Estimated kernel-density distribution of log_2 counts per million reads (CPM) of miRNAs grouped by 3′ RNA nucleotide, with (top) or without (bottom) 3′ labeling.

B, Mean fraction and standard deviation of alignment coverage to miRNA 3′ ends grouped by 3′ rR (red) and 3′ rY (blue), with (top) or without (bottom).

Input RNA 3′ labeling for optimal 3′ precision

Previous structure/function studies demonstrated that BoMoC side-chain substitutions can increase manganese-stimulated 3′ tailing of ssRNA (Pimentel et al. 2022). Here we compared previously generated BoMoC variants with side-chain substitutions W403A, G415A, F753A, or I770A (Pimentel et al. 2022), and a previously unreported double-mutant W403AF753A, for 3′ labeling of input RNA in the first step of OTTR. OTTR cDNA libraries made by BoMoC W403A, I770A, F753A or W403AF753A had significantly higher 3′ precision (one-sided student's *t*-test p-value of 2.69×10^{-2}, 4.06×10^{-4}, 4.37×10^{-11}, and 5.52×10^{-21}, respectively) and lower CV compared to cDNA libraries made using original BoMoC (BoMoC WT) for the 3′-labeling step (**Fig. 2.3A**). Libraries made using BoMoC F753A or I770A for the 3′-labeling step had the lowest library-wide CV, and libraries made with BoMoC F753A or W403AF753A had the highest 3′ precision (**Fig. 2.3A**). Hereafter, unless specified otherwise, we used BoMoC F753A or W403AF753A for the 3′-labeling step (**Fig. 2.1**, changes in red). BoMoC F753A was previously introduced as it reduced 3′ bias when used for 3′ labeling in ribosome profiling library synthesis by OTTR (Ferguson et al. 2023, Li et al. 2022, Mestre-Fos et al. 2024).

We also investigated how the buffer used for working-stock enzyme storage at -20 °C impacted enzyme activity after prolonged periods of storage. We created working stocks by 5-fold dilution of a long-term -80 °C stock into a variety of different conditions and tested the activity of these working stocks. While many conditions did not make a difference in 3′-labeling or cDNA synthesis activity when assayed immediately after dilution, differences emerged after months of storage time that could be detected by assays of 3′-labeling activity directly or in OTTR cDNA libraries. We found that the following improved stability of BoMoC 3′-labeling activity during 6 months of storage at -20 °C (data not shown): mildly acidic (pH 6 – 6.5) buffers such as bis-tris (25 mM) or arginine-HCl (0.5 M); either increasing the concentration of dithiothreitol (DTT) (5 – 15 mM) or replacing DTT with a low concentration of tris(2-carboxyethyl)phosphine (TCEP) (0.2 mM); and replacing KCl (800 mM) for a mix of KCl (200 mM) and $(NH_4)_2SO_4$ (400 mM). A correlation emerged between a drop-off in 3′-labeling activity and in reducing agent. Maintaining the presence of reduced DTT with lowered pH or higher initial DTT concentration, or using a low concentration of pH-stable TCEP instead, were all effective for reducing loss of 3′-labeling activity. A low concentration of TCEP in neutral storage buffer was optimal (**Fig. 2.3B, Table 2.1**).

Given the results above, and a concern that high DTT concentration could inhibit the phosphatase activity of the rSAP enzyme used after 3′ labeling (**Fig. 2.1**), we settled on an updated -20 °C storage diluent with neutral pH and low TCEP: 25 mM tris-HCl pH 7.5, 200 mM KCl, 400 mM $(NH_4)_2SO_4$, 0.2 mM TCEP, 50% glycerol (**Fig. 2.1**, changes in red). We also detected that over time a second buffer was sensitive to prolonged storage. After multiple free-thaw cycles within a 1 – 3 month time frame, the manganese in 3′ labeling Buffer 1B (**Fig. 2.1**) could form an oxidized yellow precipitate, which correlated with reduced 3′-labeling activity and thus reduced 3′ precision. This oxidation was defeated by exploiting prior knowledge (Hem 1963) to guide testing of replacement Buffer 1B, resulting in adoption of Buffer 1B2 with 10 mM sodium acetate at pH 5.5, 28 mM $MnCl_2$, 28 mM $(NH_4)_2SO_4$, and 3.5 mM ddATP (**Fig. 2.1**, changes in red).

With the 3′-labeling step optimized by BoMoC mutagenesis and improved storage stability, we assessed whether 3′ precision still required the 30-minute ddGTP supplementation after 90 minutes of 3′ labeling with ddATP (**Fig. 2.1**). This ddGTP supplementation suppressed the over-capture of a small fraction of RNAs that did not 3′ label with ddATP under earlier OTTR conditions (Upton et al. 2021, Ferguson et al. 2023). Capture of 3′ddG-labeled input RNA obliged the +1 T primer duplex to be spiked with a lower concentration of +1 C primer duplex, where the 3′T nucleotide was changed to 3′C, a mutation that would compromise Read2 sequencing (**Table 2.2,** +T.v2.Cy5 and +C.v2.Cy5). Also, if commercially available indexing primers that included the full-length Read2 sequence were used for library PCR, the cDNAs captured by the +1 C primer duplex would be under-amplified (**Extended Data 2.2)**. While these sequencing and amplification concerns could be readily overcome by adding an extra primer-duplex base pair (**Table 2.2**, see updated universal primers for +1 Y duplex), we hoped to eliminate the ddGTP supplementation step entirely.

To reassess 3′-labeling efficiency, we conducted time-course experiments under various conditions. While 30 minutes of 3′ labeling with ddATP by BoMoC F753A gave significantly worse 3′ precision than the standard OTTR protocol of 90 minutes of 3′ labeling with ddATP by BoMoC F753A followed by a 30-minute ddGTP supplementation, both 60 and 120 minutes of 3′ labeling with ddATP by BoMoC F753A were not significantly different for 3′ precision (**Fig. 2.3C**, orange bars; **Table 2.1**). For BoMoC W403AF753A, we noted a significant improvement for ddA-only 3′ labeling at both 60 and 120 minutes of labeling compared to the standard OTTR protocol (**Fig. 2.3C**, blue bars; **Table 2.1**). Moreover, we found the use of BoMoC F753A or W403AF753A for 120 minutes of 3′ labeling with only ddATP improved CV (0.568 and 0.597, respectively) compared to the inclusion of a ddGTP-supplementation step (0.766) (**Fig. 2.3C**; **Table 2.1**). In fact, all libraries synthesized with BoMoC F753A or W403AF753A using just ddATP showed markedly improved CV (0.568 – 0.647 or 0.597 – 0.646, respectively, **Table 2.1**).

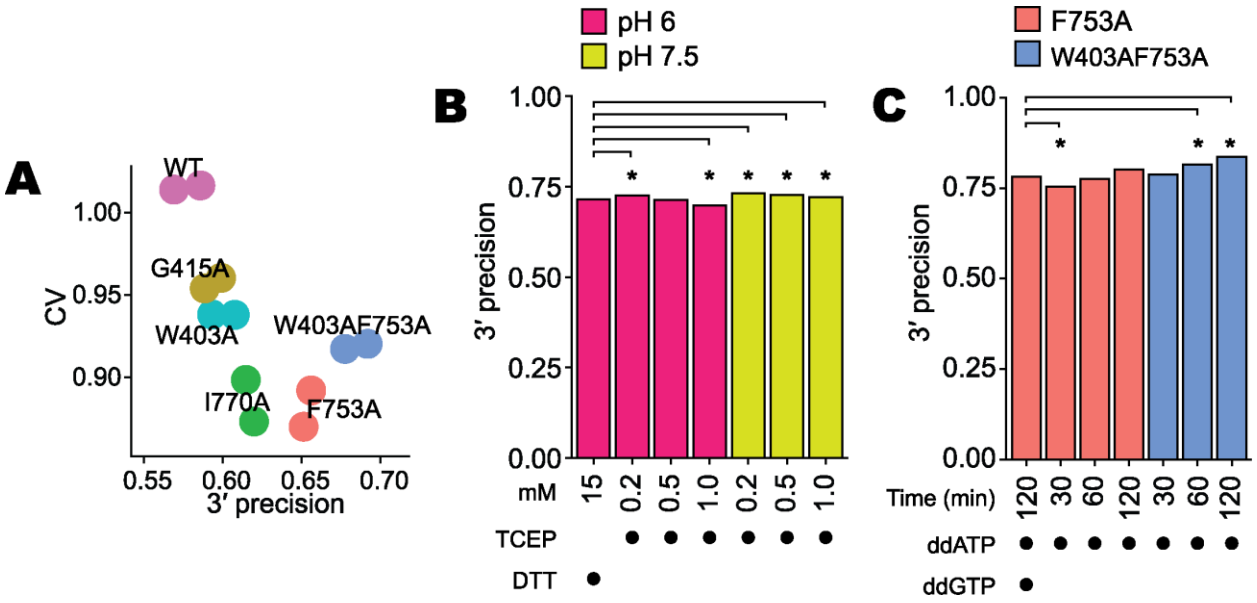
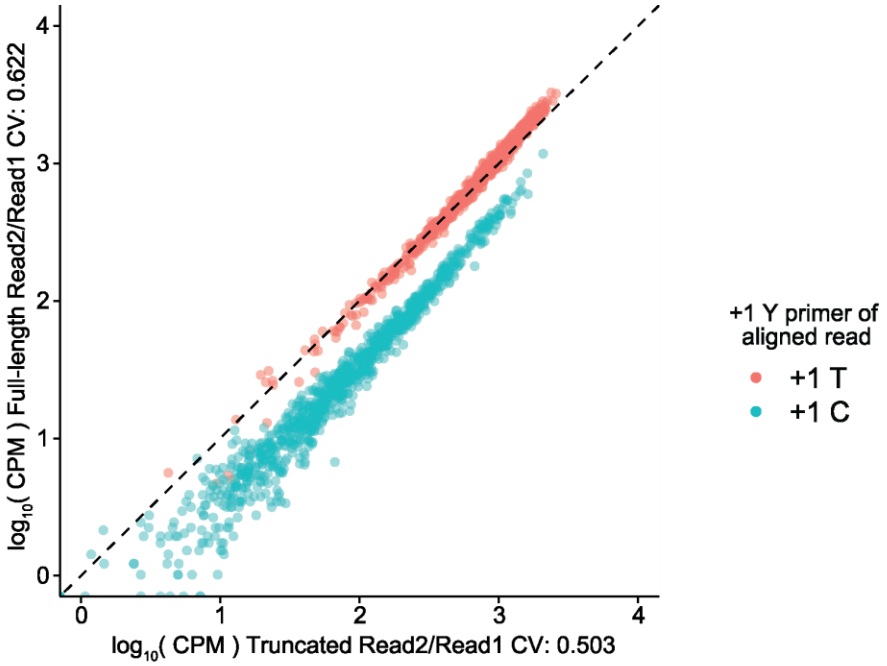


Figure 2.3: Optimization of 3′ labeling via enzyme and reaction buffers.

A, As in **Figure 2.2B**, using BoMoC WT or variant enzymes. Replicate cDNA libraries have the same color.

B, Comparison of 3′ precision under different BoMoC working-stock enzyme storage buffer conditions. Significant differences in 3′ precision (denoted by asterisks) were determined by a paired, two-sided student's *t*-test for each miRNA, with p-values adjusted by Bonferroni correction. Each library was benchmarked against a library prepared with enzyme diluted in 15 mM DTT in pH 6.0 diluent buffer.

C, Comparison of 3′ precision using different BoMoC proteins, incubation times, and ddNTP(s) for 3′ labeling. Significant differences were measured as described in **(B**).



Truncated Read2/Read1 indexing primer pair example
CAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTG
AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGAC

Full-length Read2/Read1 indexing primer pair example
CAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTG**TGCTCTTCCGATC*T**
AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGAC**GCTCTTCCGATC*T**

Extended Data 2.2: Under-amplification of cDNA initiated using primer duplex +1 C using some primer sets for library PCR.

Comparison of miRNA counts per million reads (CPM) based on how the input RNA was captured, *i.e.*, either by +1 T primer duplex or +1 C primer duplex (**Table 2.3**, +1T.v2.Cy5 and +C.v2.Cy5). Replicate cDNA libraries were split and PCR-amplified by either full-length (n=2) or truncated (n=2) Read1 and Read2 indexing primers (examples below, with the portions of the Read1 and Read2 sequences excluded from the truncated primer set shown in bold). The “[i7]” and “[i5]” indicate the index sequences (not shown).

Precision of sequence capture at input RNA 5′ ends

Following cDNA synthesis across an input-template molecule, the cDNA is extended by a +1 G due to both a deoxyguanosine triphosphate (dGTP) skew in the dNTP concentration used for the RT reaction and an inherent BoMoC preference for NTA using deoxypurine triphosphates (dRTP) (Upton et al. 2021). The +1 G directs the second template jump to the 3′rC of the 3′ adapter template rather than another 3′ddR-labeled input template (**Fig. 2.1**). However, short cDNA products that appear to lack an insert, *i.e.,* adapter-dimer cDNAs, are generated by a single template-jump from the +1 Y primer duplex to the 3′rC adapter template. These products appear with a heterogeneous electrophoretic mobility during cDNA size selection (see below and **Extended Data 2.3A**). Sanger sequencing of these short cDNAs cloned into a plasmid vector (**Table 2.3**) revealed that cDNAs with 3′ truncations of adapter template sequence were common, suggesting that the six nucleotides of RNA at the 3′ end of the original 3′ adapter template (Upton et al. 2021) were chemically labile and/or favored internal initiation. Also, in high-throughput sequencing of miRXplore cDNA libraries generated by both OTTR template-jumps, apparent miRNA one-nucleotide 5′ truncations could be detected, suggestive of 5′ imprecision in sequence capture (**Fig. 2.4A**, **ii**).

To reduce 5′ imprecision and also introduce a UMI, we evaluated the performance of different 3′ adapter-template sequences with specific and degenerate ribonucleotides between the 3′rC and Read1 sequence (see **Fig. 2.1** for schematic). We hypothesized that a 3′ penultimate rR would discourage internal initiation whereas a 3′ penultimate rY would encourage it (**Fig. 2.4A, ii** versus **iii**). We first assessed 5′ precision using 3′ adapter-template sequences in the same format as the original 3′ adapter template, comprised mostly of DNA, with RNA in the 3′-most six nucleotides, and found significant improvement in 5′ precision for all tested UMI-containing 3′ adapter templates (**Fig. 2.4B; Table 2.4**). As predicted, 5′ precision was greatest using 3′ adapter templates with rR at the 3′-penultimate position (**Fig. 2.4B; Table 2.4**).

We next tested how 5′ precision was affected by adapter template 3′-terminal ribonucleotides versus deoxyribonucleotides. We replaced some or all of the 3′-terminal RNA with DNA, including a DNA-only 3′ adapter template with 3′ dideoxycytidine (ddC) (hereafter 3′ddC adapter template). Independent of adapter template sequence, the replacement of RNA and inclusion of a 3′-terminal ddC improved 5′ precision (**Fig. 2.4C; Table 2.4**). Strategically, the 3′ddC also prevented an adapter template from priming unwanted side-reaction synthesis, a benefit that was detectable in OTTR cDNA products when the primer duplex and adapter template were labeled with distinct fluorophores (see below and **Extended Data 2.3A**). Of note, we did not detect a preference for jumping to DNA-only 3′ adapter template compared to adapter templates with six 3′ ribonucleotides when assessing OTTR cDNA libraries produced using mixed populations of adapter templates, for example, when adapter templates with six 3′ ribonucleotides and the sequence 3′rCrA were combined in a 1:2 ratio with the DNA-only 3′ddCG adapter templates (data not shown). Based on these findings, we adopted the use of a 3′ adapter template composed entirely of DNA with the terminal sequence

3′ddCR. This updated 3′ adapter template was included in the replacement of Buffer 4B with Buffer 4B2 (**Fig. 2.1,** changes in red).

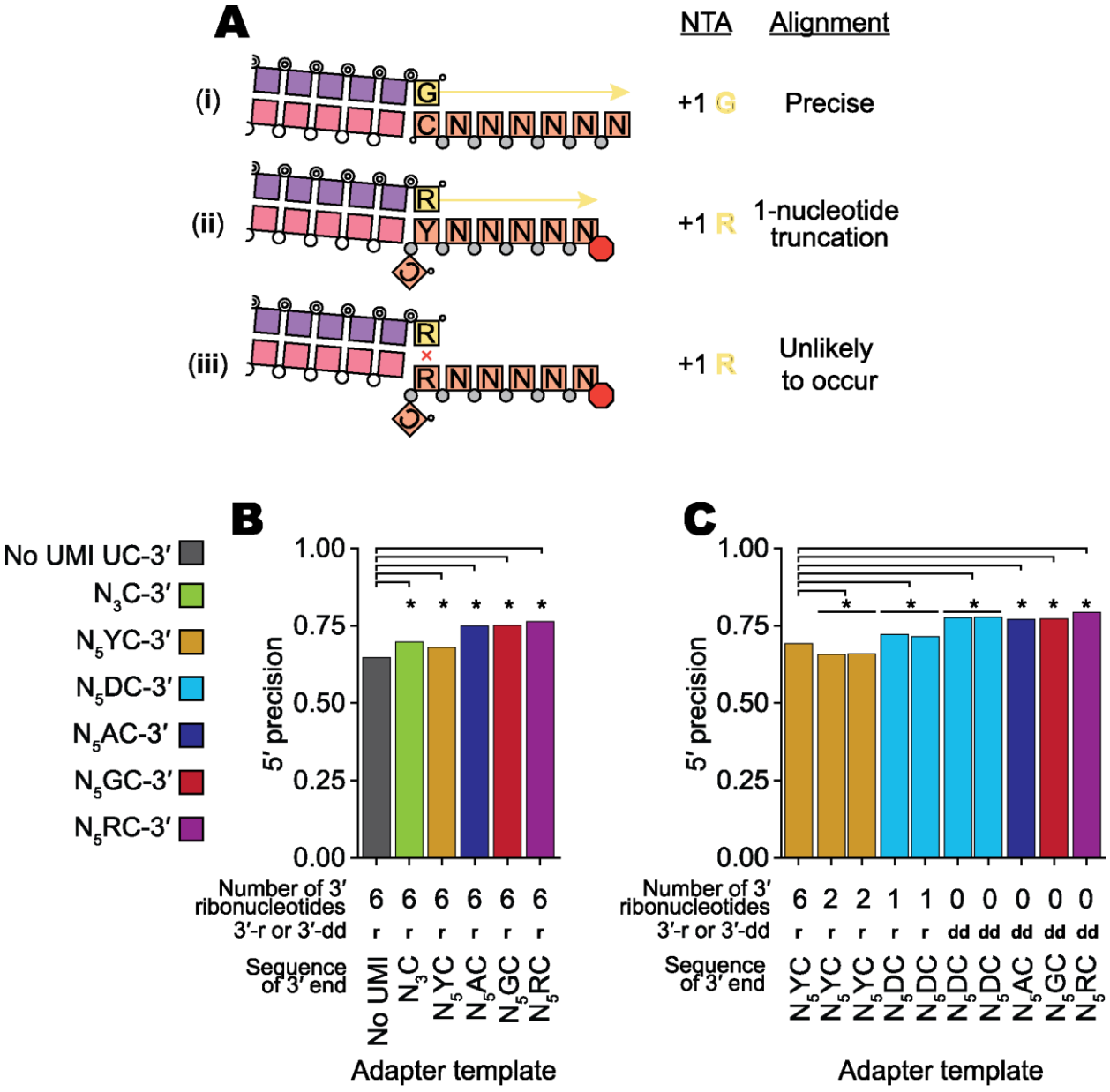


Figure 2.4: Evaluation of imprecision of end-to-end sequence capture at RNA 5′ ends.

A, Mechanisms, numbered by Roman numeral, of precise and imprecise template jumping to the 3′rC adapter template. Column one illustrates two possible and one unlikely specificity of sequence-junction formation following the color legend of **Figures 1** and **2A**. Column two defines the NTA that would prime 3′rC adapter-template capture. Column three defines whether the mapped template 5′-end was precise.

B,C, Comparison of mean fraction of alignments with 5′ precision for libraries prepared with different 3′rC adapter template designs (**Table 2.5**). Significant differences in 5′ precision (denoted by asterisk) were determined by a paired, two-sided student's *t*-test for each miRNA, with p-values adjusted by Bonferroni correction. See **Table 2.4**.

Lower limits for amount of input RNA

Optimization trials described previously (Upton et al. 2021) and above used 0.5 – 10 ng of miRXplore miRNA per OTTR reaction. To investigate OTTR performance with lower input, we titrated the total amount of miRNA pool across a range of 4 – 500 pg per 20 µL reaction. In the initial comparison, BoMoC F753A was used for 90 minutes of 3′ labeling with ddATP only and 30 additional minutes of 3′ labeling after supplementation with ddGTP. The OTTR reaction used primer duplexes with +1 Y at a 4:1 ratio of +1 T to +1 C (**Table 2.2**, +T.v2 and +C.v2 annealed to RNA.v2), with the primer DNA strands bearing either 5′ Cyanine5 (Cy5) or 5′ near-infrared (IRD800) dye for fluorescent cDNA detection, and c3t_CY5N_6RNA 3′rC adapter template (**Table 2.4**), combined with BoMoC WT for cDNA synthesis. Detection sensitivity for cDNA resolved by denaturing (d) urea polyacrylamide gel electrophoresis (PAGE) was higher using IRD800 than Cy5 (**Fig. 2.5A**), and it was also possible to distinguish 5′-IRD800-labeled cDNAs from 5′-Cy5-labeled adapter template and its side-reaction products, which can migrate by dPAGE as either single-stranded or reannealed adapter oligonucleotides at a rate faster than the full-length, insert-containing cDNAs (**Extended Data 2.3A,** purple dash, Cy5 emission). There was little variation in library-wide CV across the input titration range using either IRD800 or Cy5 to detect cDNA for gel purification (**Fig. 2.5B**). To capture all cDNA products, including cDNA from any nucleic acids contaminating the enzyme preparations, cDNA size selection encompassed an input nucleic acid size-range of up to 200 nucleotides (**Fig. 2.5A**, black vertical bar).

While library-wide CV was consistent across the input titration, the percentage of library with useful sequencing reads was not. Non-miRNA reads increased as input decreased (**Fig. 2.5C**). Most non-miRNA reads mapped either to *E. coli* nucleic acids or BoMoC expression plasmid which would be brought to the reaction by purified BoMoC (**Fig. 2.5C**, **Extended Data 2.3B**). Because the same reaction conditions were used across the input titration, including a fixed amount of each enzyme, the relative representation of bacterial contaminants increased with the decrease in input template. We conclude that detection sensitivity can be limited by the competing capture of nucleic acid impurities in the enzyme preparations, despite rigorous discrimination against this in development of the original purification (Upton et al. 2021).

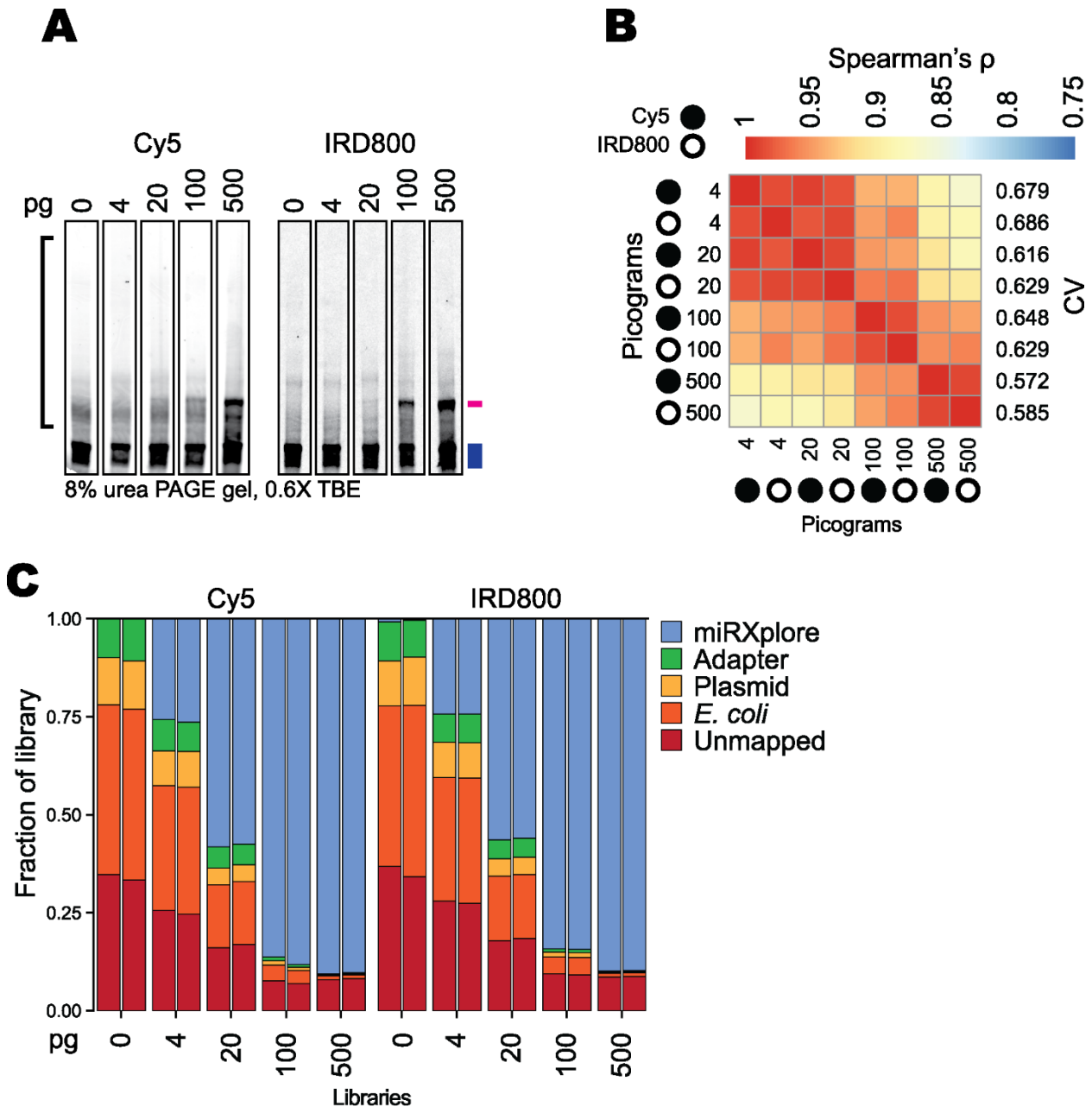


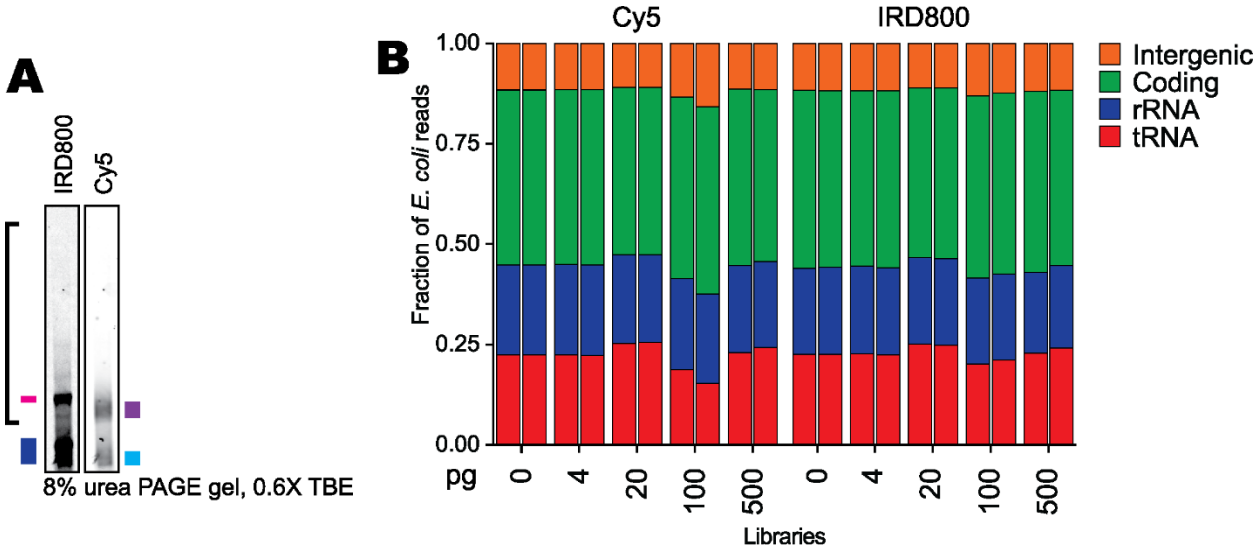
Figure 2.5: OTTR performance across different amounts of input RNA

A, Representative replicates of OTTR cDNA resolved by 8% dPAGE and detected using the 5′ Cy5 or IRD800 primer fluorophore, with miRXplore miRNA input in pg and 5′ primer fluorophore denoted above. cDNA product size-selection range is denoted by the left open bracket (black). On the right, adapter-dimer (blue bar) and miRXplore OTTR (pink dash) cDNA products are denoted.

B, Correlation matrix of miRNA read counts from cDNA libraries produced using a titration from 4 to 500 pg total miRXplore RNA. Replicate Cy5 and IRD800 libraries for each miRNA input amount were averaged by their counts per milllion reads (CPM) and

compared by Spearman's correlation coefficient (ρ). miRXplore miRNA input in pg, 5′ primer fluorophore, and library-wide CV are denoted.

C, For each library, bar plots indicate the fraction of reads that mapped to miRNA, OTTR adapter sequences, BoMoC expression plasmid, or *E. coli* genome, or were unmapped. Individual replicate libraries are presented in pairs.



Extended Data 2.3: Gel migration and sequencing of unwanted OTTR products.

A, A single OTTR cDNA library constructed with a 5′ IRD800 primer fluorophore (from lane 10, **Fig. 2.5A**) was resolved by 8% dPAGE and imaged for both IRD800 and Cy5 emission. cDNA library product size-selection range is denoted by the left open bracket (black). On the left, adapter-dimer (blue bar) and miRXplore OTTR (pink dash) cDNA products from primer with IRD800 emission signal are indicated. On the right, heterogenous migration of products with Cy5 signal (purple and light blue) is evident in imaging of 5′ Cy5 3′rC adapter template (**Table 2.5**, c3t_CY5N_6RNA).

B, Subcategories of read mapping for the *E. coli* nucleic acid category of libraries in **Figure 2.5C**.

To increase the productive sequencing of low-input samples, we tested changes to the 3-step BoMoC purification regimen. Original purifications used Rosetta2 (DE3) pLys cells, which over-express seven rare *E. coli* tRNAs from a plasmid (Upton et al. 2021). BoMoC protein production was induced overnight at 16 °C, followed by cell lysis in the presence of benzonase to degrade DNA and RNA, lysate clarification, and chromatography using nickel, heparin, and size-exclusion (SEC) in series (**Extended Data 2.4**) (Upton et al. 2021). We screened an initial panel of enzyme purification changes by production of sequencing libraries using 20 pg of miRXplore as OTTR input

and size selected for cDNA inserts from ~10 – 200 nucleotides as above. Comparison of two different enzyme purification strategies, starting from the same bacterial cell culture, revealed a change in the contaminating *E. coli* nucleic acids when polyethyleneimine (PEI) was used to clarify the cell lysate prior to chromatography (**Fig. 2.6A**, compare prep 2 that included a PEI precipitation step to prep 1 that did not). Sequencing reads that mapped to the BoMoC expression plasmid were greatly reduced, as were other *E. coli* nucleic acids, but full-length tRNAs remained (**Fig. 2.6B-C**). The level of tRNA contamination varied from purification to purification (data not shown), which we speculate resulted from slight differences in cell growth or lysis conditions in Rosetta2 (DE3) pLys cells.

To quantify the extent of nucleic acid contamination across different enzyme purifications, we established an approach to estimate the minimum miRXplore input needed for ~90% of mapped sequencing reads to map to miRXplore versus ~10% to *E. coli* or BoMoC expression plasmid. OTTR libraries with inputs of 20 or 500 pg of miRNA were produced, sequenced, and mapped to reference sequences in parallel. Purifications compared the use of the Rosetta2 strain to an *E. coli* expression strain that does not over-express tRNAs, T7 Express lysY/Iq. Purifications also compared inclusion or omission of benzonase in the cell lysis buffer, lysate clarification with PEI, and the final SEC step. Comparison of read mapping using the 20 ng input samples revealed that the change of expression strain and the use of PEI combined were sufficient to nearly eliminate contaminating bacterial nucleic acids, with some additional benefit when SEC was included (**Fig. 2.6D**). Purifications of BoMoC F753A and BoMoC WT from T7 Express cells lysed in the presence of benzonase, and with cell lysate clarified by PEI precipitation (preps 8 and 9, respectively, in **Fig. 2.6D**), gave the lowest level of contaminating nucleic acid reads, which showed reduced predominance of *E. coli* tRNA relative to *E. coli* ribosomal RNA (rRNA) and coding- or intergenic-region genomic sequences (**Fig. 2.6E**).

We used OTTR cDNA libraries produced from a range of miRXplore amounts to extrapolate the input requirement for 90% or more of reads to map to input templates instead of contaminating bacterial nucleic acids. From the input RNA titration series described above, performed using BoMoC purified with the original protocol, the input required for 90% miRXplore versus bacterial nucleic acid reads was 52.2 pg (**Fig. 2.6F**, purple data points). In comparison, using the newly optimized expression and purification conditions (**Fig. 2.6D**, preps 8 and 9), the input required for 90% miRXplore versus bacterial nucleic acid reads was only 2.80 pg (**Fig. 2.6F**, orange data points). We confirmed this extrapolated lower limit on input amount by generating OTTR cDNA libraries using a titration of miRXplore input from 0.2 – 500 pg with the same purified proteins (**Fig. 2.6G**). We conclude the updated BoMoC purification protocol substantially reduces the co-purification of nucleic acid, enabling a greater chance for success in lower-input OTTR cDNA libraires (**Extended Data 2.4**, changes in red).

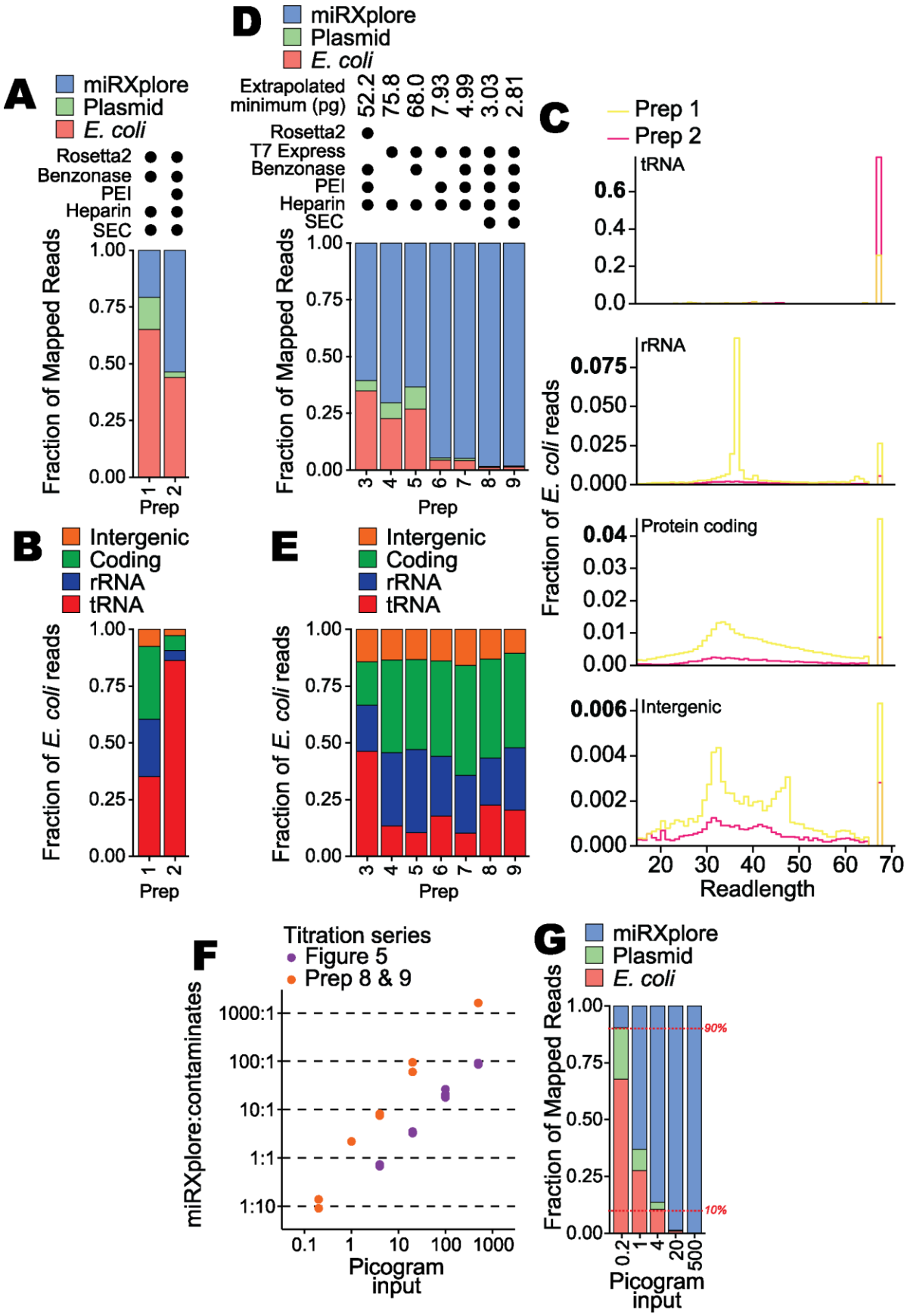

A
miRXplore
Plasmid
E. coli
Rosetta2
Benzonase
PEI
Heparin
SEC
Fraction of Mapped Reads
Prep
B
Intergenic
Coding
rRNA
tRNA
Fraction of E. coli reads
Prep
D
Extrapolated minimum (pg)
52.2
75.8
68.0
7.93
4.99
3.03
2.81
T7 Express
E
C
Prep 1
Prep 2
tRNA
rRNA
Protein coding
Intergenic
Readlength
F
Titration series
Figure 5
Prep 8 & 9
miRXplore:contaminates
1000:1
100:1
10:1
1:1
1:10
Picogram input
G
90%
10%

Figure 2.6: Improvements in BoMoC protein purification to limit co-purification of *E. coli* nucleic acids.

A, Composition bar plots of mapped reads for miRXplore cDNA libraries made using 20 pg input RNA and different protein purifications (preps). Variations to protein purification (**Extended Data 2.4**) are described using filled circles to indicate the presence of a variable *E. coli* strain or purification step in each prep. Prep 1 and 2 were BoMoC W403AF753A purified after splitting the same cell lysate.

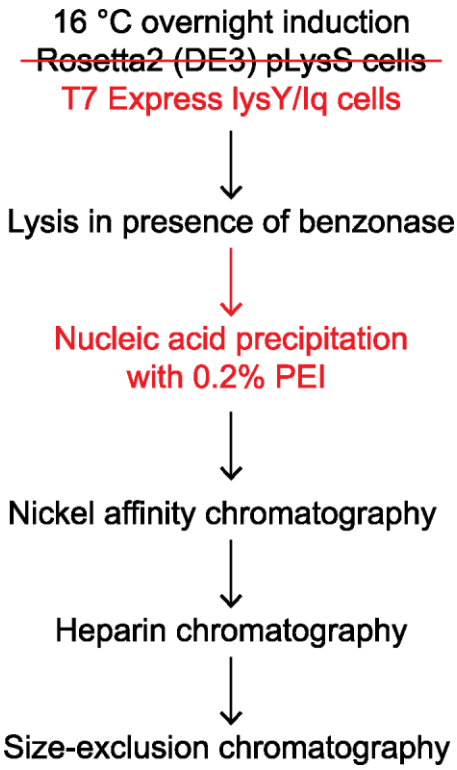
B, Subcategories of read mapping for the *E. coli* nucleic acid category in (**A**).

C, Read length distribution plots of for the subcategories of read mappings in (**B**). Note that the y-axis for each plot has a different scale.

D,E, Composition bar plots as described for (**A-B**) for cDNA libraries from 20 pg miRXplore input. In addition to 20 pg miRXplore RNA input, a 500 pg miRXplore RNA input cDNA library was produced; the pair were used to extrapolate minimum input that would recover 9:1 miRNA:*E. coli* and expression-plasmid reads, with that amount indicated at top. Preps 3 and 8 were BoMoC F753A; preps 4, 5, 6, and 7 were BoMoC W403AF753A; and prep 9 was BoMoC WT.

F, Titration results for data from **Figure 2.5** and additional results for cDNA libraries prepared using preps 8 and 9 for 3′ labeling and cDNA synthesis, respectively. The y-axis shows the read count ratio of miRXplore:contaminates (*i.e., E. coli* and expression plasmid sequences). Both axes are on a $\log_{10}$ scale.

G, Composition bar plots as described for (**A**), here for libraries constructed with 0.2, 1, 4, 20, and 500 pg of miRXplore RNA and using preps 8 and 9 for 3′ labeling and cDNA synthesis, respectively. One representative replicate from the data in (**F**) was used for the bar plot. Red horizontal dashed lines (10% and 90%) are included as visual aids.



Extended Data 2.4: Flowchart describing BoMoC purification steps.

Modifications introduced in this work are added in red text, and a red line is drawn through abandoned steps.

Materials and methods

BoMoC enzymes were purified largely as described previously (Upton et al. 2021, Ferguson et al. 2023) using nickel affinity chromatography, heparin chromatography, and size-exclusion chromatography, unless specified. Initial and improved protein purification conditions are compared in **Extended Data 4**. BoMoC has a C-terminal 6x-histidine tag for purification and N-terminal maltose binding protein fusion for solubility. Strains were Rosetta2 (DE3) pLysS (Novagen, 71403), which possesses a plasmid to over-express rare *E. coli* tRNAs, or T7 Express Lys/IQ (New England Biolabs, C3013I). Cells were transformed with an expression plasmid (*e.g.* Addgene plasmid 186461, 185710, or 185713) and cultured overnight in 100 mL of 2X yeast extract tryptone (YT) media supplemented with appropriate antibiotics. The next day, cultured cells were diluted to OD_{600} of 0.05 and incubated in a 37 °C orbital shaker until the OD_{600} was 0.5 – 0.6. The culture flasks were then transferred to a pre-chilled 16 °C orbital shaker and left to chill for 25 – 30 minutes. The OD_{600} was monitored until it reached 0.7 – 0.8, at which point isopropyl ß-D-1-thiogalactopyranoside was added to a final concentration of 0.5 mM to induce protein expression. After 12 – 16 hours, cells were harvested by centrifugation for 20 minutes at 3600 x g at 4 °C. The media was aspirated from the cell culture pellet and replaced with 15 mL of lysis buffer (20 mM Tris-HCl pH 7.4, 1 M NaCl,

10% glycerol, 1 mM β-mercaptoethanol, 1 µg/ml pepstatin A, 1 µg/ml leupeptin, 0.5 mM phenylmethylsulfonyl fluoride) per L of culture. After resuspension in lysis buffer, the cell lysate was frozen in liquid nitrogen and stored at -80 °C.

Frozen cell lysate was thawed in a room-temperature water bath before supplementing with $MgCl_2$ to 0.5 mM. If specified, a volume of benzonase (Millipore) was added corresponding to a 1:1000 dilution. The thawed lysate was transferred to ice and sonicated for 3.5 minutes with 10 seconds of sonication separated by 10 seconds of rest. Insoluble material was pelleted by centrifugation using a SS34 rotor at 15,000 RPM for 30 minutes at 4 °C. Supernatant was decanted and recentrifuged at 5,000 x g for 10 minutes at 4 °C to remove any additional insoluble material from the cell lysate supernatant. For preps where PEI precipitation was included, the cell lysate was decanted into a bottle with a stir bar and set to stir rapidly without foam or excess bubbles. Prior to this, a 10% w/v PEI solution was made fresh, first by stirring 10 g PEI in 40 mL water while adding concentrated HCl dropwise until the pH was ~7.2. Once cooled, the pH was adjusted to pH 7.0 – 7.4 and the volume was adjusted with additional water to to 50 mL total. The neutralized 10% w/v PEI solution was added dropwise to the cell lysate until a final concentration of 0.2% w/v PEI was reached. Nucleic acids were pelleted by centrifugation using a SS34 rotor at 15,000 RPM for 30 minutes at 4 °C, and the supernatant was decanted as clarified cell lysate.

All chromatography steps were at room temperature. For nickel affinity chromatography, clarified lysate was loaded to two 5 mL HisTrap FF Crude columns (Cytiva) connected in series. Unbound protein was removed by 5 – 10 column volumes of Nickel A buffer (20 mM Tris-HCl pH 7.4, 1 M KCl, 20 mM imidazole, 10% glycerol, 1 mM β-mercaptoethanol). Bound protein was eluted with five column volumes Nickel B buffer (20 mM Tris-HCl pH 7.4, 1 M KCl, 400 mM imidazole, 10% glycerol, 1 mM β-mercaptoethanol). Fractions were measured for protein content a A_{280}, and fractions with protein were pooled and filtered with a 0.22 µm polyethersulfone syringe filter. Next, pooled eluent was desalted to the equivalent of 20% (*i.e.,* ~400 mM KCl) Heparin B buffer (25 mM HEPES-KOH pH 7.5, 2 M KCl, 10% glycerol, 1 mM DTT) and loaded to a 5 mL Heparin HP column (Cytiva). Unbounded protein was removed with 20% Heparin B buffer. Bound protein was eluted in 15 column volumes with a gradient of 20% – 100% Heparin B buffer. Fractions were measured for protein content a A_{280}, and those with protein were pooled and concentrated to 4 – 4.5 mL before filtered with a 0.22 µm polyethersulfone syringe filter. A SEC HiPrep 16/60 Sephacryl S300 column (Cytiva) equilibrated in SEC buffer (25 mM HEPES-KOH pH 7.5, 800 mM KCl, 10% glycerol, 5 mM DTT) was loaded with sample followed by one column volume of SEC buffer. Fractions in the size range of monomeric protein were measured for protein content at A_{280}, and those with protein were pooled and concentrated to 8 mg/mL (50 µM) in SEC buffer using an Amicon centrifugal filter unit.

Aliquots were made at a volume no greater than 100 µL and were snap-frozen in liquid nitrogen and stored long-term at -80 °C. To make a working stock, the 100 µL 50 µM aliquot was rapidly thawed by directly transferring the tube from the -80 °C to a beaker of pre-warmed 37 °C water for <30 seconds. Once thawed, it was diluted to 10 µM by combining with 400 µL pre-chilled 4 °C storage diluent buffer (optimally 25 mM

Tris-HCl pH 7.5, 200 mM KCl, 400 mM $(NH_4)_2SO_4$, 0.2 mM TCEP, 50% glycerol; see main text for variations) and transferred to ice. The diluted enzyme stock was then carefully mixed by pipetting 50% of the volume ~50 times on ice.

OTTR cDNA library synthesis, cDNA purification, and cDNA size-selection

All oligonucleotides used were synthesized by Integrated DNA Technologies, with RNase-free high performance liquid chromatography purification included for oligonucleotides used in cDNA library synthesis. As described previously (Upton et al. 2021, Ferguson et al. 2023), to make Buffer 4B or Buffer 4B variants, primer-duplex DNA and RNA strands were first heated to denature structure at 70 °C for 5 minutes before annealing by decreasing the temperature to 4 °C with a -0.2 °C/second ramp. Annealed duplexes were made at 50 µM in water, diluted to 3.6 µM with additional water, and combined 1:1 with 7.2 µM 3′rC adapter template or 3′ddC adapter template also in water, resulting in a final Buffer 4B with 1.8 µM primer duplex and 3.6 µM adapter template.

All libraries in this study were synthesized using miRXplore Universal Reference Standard (Miltenyi Biotech, 130-094-407) as input RNA, or with no input template. OTTR workflow and buffers are summarized in **Figure 2.1**. Input miRXplore miRNA was diluted to 9 or 10 µL in nuclease-free water. 2 µL of Buffer 1A, 1 µL of Buffer 1B or 1B2, and 1 µL of 10 µM working-stock BoMoC (either WT or mutant) or yPAP (Thermo Scientific, 74225Z25KU) were added sequentially. After 90 minutes at 30 °C (or other time, if indicated),1 µL of Buffer 1C was added (or not, if indicated) and the reaction was either incubated for an additional 30 minutes (or not, if indicated) at 30 °C. 1 µL of Buffer 2 and 0.5 µL of rSAP (New England Biolabs, M0371S) pre-mixed with 0.5 µL 50% glycerol were added. After 15 minutes at 37 °C, 1 µL of Buffer 3 was added. After 5 minutes at 65 °C, the reaction was placed on ice. 1 µL of Buffer 4A, 1 µL of Buffer 4B or Buffer 4B2, and 1 µL of the 10 µM working-stock BoMoC were added sequentially, followed by incubation for 30 minutes at 37 °C.

After OTTR, cDNA products were recovered largely as described previously (Ferguson et al. 2023). BoMoC was inactivated by incubation at 70 °C for 5 minutes, followed by addition of 1 µL RNase A (Sigma, R6513) and 1 µL of Thermostable RNase H (New England Biolabs, M0523S) before incubation at 55 °C for 15 minutes. Then 1 µL of Protease K (New England Biolabs, P8107S) and 30 µL of stop buffer (50 mM tris-HCl pH 8.0, 20 mM EDTA, 0.1% SDS) were added. After 15 minutes at 55 °C, the reaction was incubated at 95 °C for 5 minutes. Zymo Oligo Clean and Concentrator-5 columns were used for cDNA purification with elution in 6 µL nuclease-free water. 6 µL of 2X formamide loading dye (FLD, 95% formamide, 5 mM ethylenediaminetetraacetic acid (EDTA) pH 8.0, 0.05% (w/v) bromophenol blue, and 0.005% (w/v) xylene cyanol) was added. cDNA was denatured at 98 °C for 5 minutes and snap-cooled on ice prior to electrophoresis resolution on a 8% 19:1 acryl:bis-acryl 7 M urea 0.6X tris-borate-EDTA (TBE) dPAGE gel pre-run and run in 0.6X TBE buffer until the xylene cyanol dye-front was near the end of the gel. The gel was imaged to detect Cy5 and/or IRD800 using an

Amersham Typhoon Trio (Cytiva) and the image was printed at 100% size of the gel to guide cDNA size-selection. Excised gel fragments were crushed against the sides of a 1.5 mL tube and cDNA was eluted into 400 – 500 µL of DNA elution buffer (300 mM NaCl, 10 mM Tris-HCl pH 8.0, and 1 mM EDTA) with a 70 °C incubation for 1 hour. cDNA was ethanol precipitated and resuspended in 30 µL nuclease-free water. cDNA was quantified by qPCR as described (Ferguson et al. 2023, McGlincy and Ingolia 2017) using iTaq™ Universal SYBR® Green Supermix (Bio-Rad, 1725120).

OTTR library PCR amplification, purification, quantification, and pooling

For sequencing, 1 – 30 µL of cDNA was amplified by 6 to 12 PCR cycles using Illumina indexing primers and Q5 polymerase (New England Biolabs, M0491). For a typical miRXplore miRNA library from 500 ng input, 4 – 6 cycles were used. For libraries from 100 ng input, 10 cycles were used. For libraries from 20 pg input, 11 – 12 cycles were used. For libraries with less than 20 pg input, 12 – 14 cycles were used. PCR products were purified by DNA precipitation, Zymo DNA Clean and Concentrate columns, or AMPure XP (A63880, Beckman Coulter) following the recommendations of the manufacturer. DNA was resolved by 8% native PAGE and eluted as described (Ferguson et al. 2023). Eluted DNA libraries were quantified by qPCR using NEBNext® Library Quant DNA Standards (New England Biolabs, E7642S) and iTaq™ Universal SYBR® Green Supermix as described (Ferguson et al. 2023). Libraries were pooled based on these results. A maximum of 30 libraries were pooled at a time, with each library indexed by unique i5 and unique i7 sequences.

Library sequencing and adaptor trimming

All libraries were sequenced on an Illumina MiniSeq instrument using 75- (Illumina, FC-420-1001) or 150- (Illumina, FC-420-1002) cycle single-end high-output reagents. Libraries were denatured following Illumina recommendations. A final concentration of 1.0 – 1.2 pM denatured library in 500 µL of HT1 hybridization buffer (Illumina, 20015892) with 5% phiX (Illumina, FC-110-3001) was loaded to the sequencer. A sequential adapter trimming pipeline using cutadapt (v3.7) was used to trim FASTQ reads and retain crucial information about the reads (Ferguson et al. 2023). In detail, the adapter trimmed in the first step corresponds to a sequence encoded in the duplexed portion of the +1 primer duplex. In the second step, X nucleotide(s) were removed from the beginning of the read, with X=1 when no UMI (**Table 2.5**, c3t_NoUMI_6RNA) was used, X=4 when the 3′CNNN adapter template (**Table 2.5**, c3t_N3C_6RNA) was used, and X=7 when all other UMI-encoding adapter templates were used. In the third step, the final nucleotide that represents the primer-duplex +1 position was removed. In both the second and third steps, the sequences removed from the read were retained in the header of a FASTQ entry. In the final step, low-quality reads were trimmed and reads shorter than 15 nucleotides were removed.

```
cat untrimmed.R1.fastq | cutadapt -a
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC - | cutadapt -u X –
rename='{id} UMI={cut_prefix}' - | cutadapt -u -1 –
rename='{id}_{comment}_PD={cut_suffix}' - | cutadapt -m 15 -q 10 - >
trimmed.fastq
```

Sequence analysis

The miRXplore reference contains 962 synthetic equimolar miRNAs ranging from 16 – 28 nucleotides in length. The 943 miRNAs of 19 – 24 nucleotides were used as a reference for read mapping. Across all libraries analyzed in this work, we calculated median read count for each miRNA; miRNAs with a median read count less than 50 were not used, restricting the analysis to 904 miRNAs. In general, we used the superset of 904 miRNAs, however for the analyses in Figures 2 and 3 investigation of 5′ and 3′ precision of end-capture, we restricted the analysis to a set of 594 miRNAs based on their inability to misalign when the first and final three nucleotides were removed before mapping. In general, we allowed alignments as short as 15 nucleotides, but for the analyses in Figures 5 and 6, alignments shorter than 17 nucleotides were excluded for mapping confidence. All libraries were trimmed as described above before a sequential alignment pipeline using bowtie (v1.0.0) was performed. Trimmed reads were first aligned to a reference of OTTR oligonucleotide adapters. Unaligned reads were next aligned to the BoMoC expression-plasmid reference then the *E. coli* BL21 DE3 genome (NCBI AM946981.2). Filtering and sorting of alignments was performed by samtools (v1.7). miRXplore miRNA single-mapping alignments to both the reliable and unreliable miRNAs were converted to BED files by bedtools bamtobed (v2.25.0).

Sanger sequencing of adapter-dimer cDNAs

Adapter-dimer-sized cDNAs were size-selected and eluted from an 8% dPAGE as described above. Eluted cDNA was amplified by PCR using Illumina indexing primers as described above, but with 20 – 30 cycles and subsequent size-selection from a 2% agarose gel. Products were cloned into pUC19 vector.

Assessment of *E. coli* nucleic acids co-purified with BoMoC

Size selection used a cDNA size range of adapters plus 10 – 200 nucleotides. Eluted cDNAs for the 500 pg input libraries were amplified by 6 – 8 PCR cycles, while cDNAs for the 20 pg input libraries were amplified by 10 – 12 PCR cycles. Libraries were enriched by native PAGE selection of ~160 – 350 bp duplexes. After sequencing and alignment as described above, the ratio of miRNA to *E. coli* and expression-plasmid reads was derived for all reads 18 nucleotides or longer. A line slope was computed from the ratio of reads mapped to miRXplore:contaminate and pg of miRXplore input.

Statistical analysis

Comparisons of the fraction of precise 5′ or 3′ alignments were made by either unpaired one-sided or paired two-sided student's *t*-tests performed only on miRNAs which had ≥30 alignments in both libraries being compared. For paired two-sided student's *t*-test, the p-values were adjusted by Bonferroni correction and the results were represented as a mean difference in the fraction of precise 5′ or 3′ alignments. A result with an adjusted p-value less than 0.05 was defined as significant. Details and results from the paired two-sided student's *t*-test were summarized in **Table 2.1** and **Table 2.5**.

Tables

Table 2.1: Results from paired, two-sided student's *t*-test of 3′ precision.

Fig.	Diluent	3′-labeling RT	3′ label	Timing (m)	CV	Mean difference	CI 95%	adjusted p-value	miRNA n	3′ precision
Fig. 2.3B	pH 6.0, 0.2 mM TCEP	F753A	ddR	90/30	0.87	1.02%	(0.84% - 1.2%)	3.98E-24	902	0.72
Fig. 2.3B	pH 6.0, 0.5 mM TCEP	F753A	ddR	90/30	0.91	-0.22%	(-0.42% - -0.03%)	1.00E+00	899	0.71
Fig. 2.3B	pH 6.0, 1.0 mM TCEP	F753A	ddR	90/30	0.86	-1.77%	(-1.97% - -1.56%)	4.38E-53	902	0.70
Fig. 2.3B	pH 7.5, 0.2 mM TCEP	F753A	ddR	90/30	0.94	1.61%	(1.4% - 1.83%)	1.26E-41	900	0.73
Fig. 2.3B	pH 7.5, 0.5 mM TCEP	F753A	ddR	90/30	0.92	1.18%	(0.96% - 1.39%)	3.14E-22	899	0.73
Fig. 2.3B	pH 7.5, 1.0 mM TCEP	F753A	ddR	90/30	1.04	0.65%	(0.43% - 0.87%)	5.85E-06	894	0.72
Fig. 2.3C	pH 7.5, 0.2 mM TCEP	F753A	ddA	30	0.65	-2.73%	(-3.73% - -1.72%)	1.18E-04	903	0.75
Fig. 2.3C	pH 7.5, 0.2 mM TCEP	F753A	ddA	60	0.63	-0.51%	(-1.51% - 0.49%)	1.00E+00	903	0.78
Fig. 2.3C	pH 7.5, 0.2 mM TCEP	F753A	ddA	120	0.57	1.98%	(1.01% - 2.94%)	6.08E-02	903	0.80
Fig. 2.3C	pH 7.5, 0.2 mM TCEP	W403A F753A	ddA	30	0.64	0.63%	(-0.39% - 1.65%)	1.00E+00	903	0.79
Fig. 2.3C	pH 7.5, 0.2 mM TCEP	W403A F753A	ddA	60	0.61	3.37%	(2.36% - 4.38%)	8.18E-08	903	0.81
Fig. 2.3C	pH 7.5, 0.2 mM TCEP	W403A F753A	ddA	120	0.59	5.51%	(4.48% - 6.54%)	2.12E-21	903	0.84

Notes: Figure panel, enzyme diluent, 3′-labeling enzyme, 3′ label nt, and timing of 3′ labeling were included in the first five columns. For 3′-labeling time, 90/30 referred to 90 minutes of 3′ labeling with ddATP by BoMoC followed by a 30-minute ddGTP-supplementation. In columns six the library wide CV was included. In column seven the mean difference in 5′ precision and a 95% confidence interval (column eight) were computed from a paired, two-sided student's *t*-test, with the p-valued adjusted (column nine) by Bonferroni correction using the number of miRNAs (n, column 10) which were evaluated. Mean fraction of 3′ precise alignments was included in column 11. For **Figure 2.3B**, all libraries were compared to a library prepared by F753A diluted in pH

6.0 15 mM DTT diluent buffer and labeled with ddRTP for 90/30 minutes. For **Figure 2.3C**, all libraries were compared to a library prepared by F753A diluted in pH 7.5 0.2 mM TCEP diluent buffer and labeled with ddRTP for 90/30 minutes.

Table 2.2: Primer duplex oligonucleotides used to capture a 3′ddA (+/- 3′ddG) labeled input RNA in OTTR.

	Name	Sequence	Mix ratio	Notes
Previously described +1 Y duplex primers	+T.v2.Cy5	/5Cy5/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	4	
	+C.v2.Cy5	/5Cy5/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCC	1	Incomplete R2 sequence. Paired-end sequencing will be inhibited. PCR amplification can be inhibited.
	RNA.v2	rGrArUrCrGrGrArArGrAmGmCmAmCmAmCmGmUmCmUmGmAmAmCmUmCmCmAmGmU/3SpC3/	5	2′-O-methyl ribonucleotides are immune to RNase H, can potentially interfere with cDNA migration during dPAGE.
Updated universal primers for +1 Y duplex	+T.v3	/5IRD800/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTT	4	Read2 Adapter trimmed from Read1 sequencing will need to be updated in the analysis pipeline.
	+C.v3	/5IRD800/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTC	1	
	RNA.v3	rArGrArUrCrGrGrArArGrAmGmCmAmCmAmCmGmUmCmUmGmAmAmCmUmCmCmAmGmU/3SpC3/	5	2′-O-methyl ribonucleotides are immune to RNase H, can potentially interfere with cDNA migration during dPAGE.
Current Buffer 4B2 primer duplex *	* +T.v2.IR	/5IRD800/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	1	Current recommendation for ddA-only labeling.
	* RNA.v2	rGrArUrCrGrGrArArGrAmGmCmAmCmAmCmGmUmCmUmGmAmAmCmUmCmCmAmGmU/3SpC3/	1	2′-O-methyl ribonucleotides are immune to RNase H, can potentially interfere with cDNA migration during dPAGE.
Updated RNase H sensitive (Buffer 4B2.1 primer duplex)	+T.v2.IR	/5IRD800/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	1	Current recommendation for ddA-only 3′ labeling.
	RNA.v2.1	rGrArUrCrGrGrArArGrAmGmCmAmCmAmCmGmUrCrUrGrArArCrUmCmCmAmGmUmCmAmC/3SpC3/	1	RNA ribonucleotides introduced between 2′-O-methyl ribonucleotides to ensure primer duplex sensitivity to RNase H.

Notes: Names, sequences, mixing ratio, and descriptions of oligonucleotides annealed together to be used as +1 primer duplexes in this study. Mix ratio described the molar proportion which duplexes were annealed.

*Current recommendation for OTTR and what is described as in Buffer 4B2.

Table 2.3: Summary of Sanger sequencing of adapter-dimer cDNAs clones.

Buffer 4B	Clone	Read2:cRead1 Junction	primer duplex	primer duplex paired	1st cDNA base	1st cDNA base paired	primer duplex dependent?	NTA dependent?	Detectable 2.by Read1 sequencing?
1Y primer duplex v2 & 3'rC no UMI adapter template	25-09	C|GAGATCGG	+1C	No	G	0	FALSE	TRUE	TRUE
	26-07	C|GAGATCGG	+1C	No	G	0	FALSE	TRUE	TRUE
	25-13	T|GAGATCGG	+1T	No	G	0	FALSE	TRUE	TRUE
	26-05	C|-AGATCGG	+1C	rC*	A	-1	FALSE	TRUE	TRUE
	27-05	C|-AGATCGG	+1C	rC*	A	-1	FALSE	TRUE	TRUE
	25-03	T|-AGATCGG	+1T	rC*	A	-1	FALSE	TRUE	TRUE
	25-04	T|-AGATCGG	+1T	rC*	A	-1	FALSE	TRUE	TRUE
	25-05	T|-AGATCGG	+1T	rC*	A	-1	FALSE	TRUE	TRUE
	25-06	T|-AGATCGG	+1T	rC*	A	-1	FALSE	TRUE	TRUE
	25-10	T|-AGATCGG	+1T	rC*	A	-1	FALSE	TRUE	TRUE
	25-14	T|-AGATCGG	+1T	rC*	A	-1	FALSE	TRUE	TRUE
	25-15	T|-AGATCGG	+1T	rC*	A	-1	FALSE	TRUE	TRUE
	26-08	T|-AGATCGG	+1T	rC*	A	-1	FALSE	TRUE	TRUE
	26-09	T|-AGATCGG	+1T	rC*	A	-1	FALSE	TRUE	TRUE
	27-06	T|-AGATCGG	+1T	rC*	A	-1	FALSE	TRUE	TRUE
	27-09	T|-AGATCGG	+1T	rC*	A	-1	FALSE	TRUE	TRUE
	28-03	T|-AGATCGG	+1T	rC*	A	-1	FALSE	TRUE	TRUE
	28-05	T|-AGATCGG	+1T	rC*	A	-1	FALSE	TRUE	TRUE
	25-01	C|------GG	+1C	rG	G	-6	TRUE	FALSE	FALSE
	26-04	C|------GG	+1C	rG	G	-6	TRUE	FALSE	FALSE
	26-06	C|------GG	+1C	rG	G	-6	TRUE	FALSE	FALSE
	27-02	C|------GG	+1C	rG	G	-6	TRUE	FALSE	FALSE
	27-04	T|------GG	+1T	rG*	G	-6	FALSE	TRUE	FALSE
	25-08	T|-------G	+1T	G*	G	-7	FALSE	TRUE	FALSE
	26-03	T|-------G	+1T	G*	G	-7	FALSE	TRUE	FALSE
	26-10	T|-------G	+1T	G*	G	-7	FALSE	TRUE	FALSE
	27-08	T|-------G	+1T	G*	G	-7	FALSE	TRUE	FALSE
	27-01	T|-------G	+1T	G*	G	-7	FALSE	TRUE	FALSE
	27-10	T|-------G	+1T	G*	G	-7	FALSE	TRUE	FALSE
	28-02	T|-------G	+1T	G*	G	-7	FALSE	TRUE	FALSE

[Table continued on next page]

	28-04	T|-------G	+1T	G*	G	-7	FALSE	TRUE	FALSE
+1 Y primer duplex v2 & 3′ddCRN5 UMI adapter template	29-01	T|GTTTTCTAGATCGG	+1T	No	G	0	FALSE	TRUE	TRUE
	30-08	C|GCGGCCCAGATCGG	+1C	No	G	0	FALSE	TRUE	TRUE
	30-03	T|--AGCTTAGATCGG	+1T	A	A	-2	TRUE	TRUE	TRUE
	29-05	T|--AACTTAGATCGG	+1T	A	A	-2	TRUE	TRUE	TRUE
	30-01	C|--GGCTAAGATCGG	+1C	G	G	-2	TRUE	TRUE	TRUE
	30-02	T|--GGGCTAGATCGG	+1T	A	G	-2	TRUE	TRUE	TRUE
	30-05	C|--GGTACAGATCGG	+1C	G	G	-2	TRUE	TRUE	TRUE
	30-10	T|--GTCATAGATCGG	+1T	A	G	-2	TRUE	TRUE	TRUE
	29-04	T|--GCGTAAGATCGG	+1T	A	G	-2	TRUE	TRUE	TRUE
	30-04	T|--GAGAAAGATCGG	+1T	A	G	-2	TRUE	TRUE	TRUE
	30-09	T|--GGGTTAGATCGG	+1T	A	G	-2	TRUE	TRUE	TRUE
	29-08	T|---GGTTAGATCGG	+1T	N	G	-3	TRUE	TRUE	TRUE

Notes: Column one described what Buffer 4B was used. The +1 Y primer duplex v2 (**Table 2.2**), 3′rC no UMI adapter template (**Table 2.5**, c3t_NoUMI_6RNA)**,** and 3′ddCRN5 (**Table 2.5**, c3t_ddCR5N) were defined below. Column two gives clone identifier. Column three gives the sequenced Read2:cRead1 junction (**Fig. 2.1**), with hyphens used to emphasizes what 3′r/ddC adapter template sequence was excluded from the junction. Column four gives the primer duplex detected. Column five gives what the primer duplex +1 nucleotide would be annealed or unpaired/mismatched (*) with the 3′r/ddC adapter template. “No” was used if no sequence of the 3′r/ddC adapter template could be attributed as either paired or mismatched with the primer duplex +1 primer duplex. Column six gives the first nucleotide polymerized by BoMoC from the primer duplex +1 primer duplex. Column seven gives the position relative to the 3′ end of the 3′r/ddC adapter template of what column six would be annealed to. Column eight provides a Boolean to describe whether the adapter dimer was dependent on a correct base-pair between the primer duplex +1 nucleotide and the 3′r/ddC adapter template. Column nine provides a Boolean to describe whether the adapter dimer was dependent on a NTA to the +1 primer duplex to template jump to the 3′r/ddC adapter template. Column 10 provides a Boolean to describe whether the adapter-dimer cDNA could be sequenced on an Illumina instrument, as assessed by the synthesis of a completed cRead2 sequence.

Table 2.4: Results from paired, two-sided student's *t*-test of 5′ precision.

Fig.	3′r/ddC adapter template sequence	3′r/ddC adapter template RNA count	3′r/ddC adapter template end	CV	Mean difference	CI 95%	adjusted p-value	miRNA n	5′ precision
Fig. 2.4B	NNNC-3′	6	3′-OH	1.07	4.92%	(4.64% - 5.21%)	8E-160	902	0.7
Fig. 2.4B	NNNNNYC-3′	6	3′-OH	0.65	3.3%	(2.99% - 3.62%)	2.8E-74	903	0.68
Fig. 2.4B	NNNNNAC-3′	6	3′-OH	0.72	10.1%	(9.56% - 10.65%)	8E-176	889	0.75
Fig. 2.4B	NNNNNGC-3′	6	3′-OH	0.73	10.45%	(9.98% - 10.91%)	2E-221	883	0.75
Fig. 2.4B	NNNNNRC-3′	6	3′-OH	0.81	11.54%	(11.04% - 12.03%)	3E-232	881	0.76
Fig. 2.4C	NNNNNYC-3′	2	3′-OH	0.66	-3.4%	(-3.56% - -3.23%)	4E-202	904	0.66
Fig. 2.4C	NNNNNYC-3′	2	3′-OH	0.65	-3.35%	(-3.51% - -3.18%)	7E-196	904	0.66
Fig. 2.4C	NNNNNDC-3′	1	3′-H	0.68	2.95%	(2.73% - 3.17%)	1E-112	904	0.72
Fig. 2.4C	NNNNNDC-3′	1	3′-H	0.69	2.24%	(2.02% - 2.46%)	1.1E-72	904	0.71
Fig. 2.4C	NNNNNDC-3′	0	3′-H	0.68	8.46%	(8.21% - 8.71%)	0	904	0.78
Fig. 2.4C	NNNNNDC-3′	0	3′-H	0.66	8.53%	(8.3% - 8.77%)	0	904	0.78
Fig. 2.4C	NNNNNAC-3′	0	3′-H	0.56	7.87%	(7.37% - 8.36%)	2E-141	904	0.77
Fig. 2.4C	NNNNNGC-3′	0	3′-H	0.61	8.06%	(7.62% - 8.49%)	1E-175	903	0.77
Fig. 2.4C	NNNNNRC-3′	0	3′-H	0.62	10.05%	(9.64% - 10.46%)	2E-249	904	0.79

Notes: Figure panel and descriptions of the 3′r/ddC adapter template (sequence, RNA composition, and whether the 3′ was unblocked, *i.e.*, 3′-OH, or blocked, *i.e.*, 3′-hydrogen (3′-H)) are described in the first four columns. See **Table 2.5** below for complete 3′r/ddC adapter template sequences. In columns five the library wide CV was included. In column six, the mean difference in 5′ precision and a 95% confidence interval (column seven) were computed from a paired, two-sided student's *t*-test, with the p-valued adjusted (column eight) by Bonferroni correction using the number of miRNAs (n, column nine) which were evaluated. Mean fraction of 5′ precise alignments was included in column 10. For **Figure 2.4B**, all libraries were compared to a library prepared with no UMI in the 3′rC adapter template, and for **Figure 2.4C**, all libraries were compared to a library prepared with 3′rC adapter template with 6-RNA ribonucleotides and a 3′-OH with the following 3′ sequence: NNNNNYC-3′.

Table 2.5: 3′r/ddC adapter template oligonucleotide sequences.

Name	Sequence	RNA ribonucleotides	UMI trimmed
c3t_NoUMI_6RNA (Upton HE. et al., 2021)	/5Cy5/ACACTCTTTCCCTACACG ACGCTCTTCCrGrArUrCrUrC	6	C
c3t_N3C_6RNA	/5Cy5/ACACTCTTTCCCTACACG ACGCTCTTCCGATrCrUrNrNrNrC	6	NNNC
c3t_CY5N_6RNA (Ferguson L. et al., 2023)	/5Cy5/ACACTCTTTCCCTACACG ACGCTCTTCCGATCTNrNrNrNrNr YrC	6	NNNNNYC
c3t_rCA5N_6RNA	/5Cy5/ACACTCTTTCCCTACACG ACGCTCTTCCGATCTNrNrNrNrNr ArC	6	NNNNNAC
c3t_CG5N_6RNA	/5Cy5/ACACTCTTTCCCTACACG ACGCTCTTCCGATCTNrNrNrNrNr GrC	6	NNNNNGC
c3t_CR5N_6RNA	/5Cy5/ACACTCTTTCCCTACACG ACGCTCTTCCGATCTNrNrNrNrNr RrC	6	NNNNNRC
c3t_CY5N_2RNA	/5Cy5/ACACTCTTTCCCTACACG ACGCTCTTCCGATCTNNNNNrYr C	2	NNNNNYC
c3t_CD5N_1RNA	/5Cy5/ACACTCTTTCCCTACACG ACGCTCTTCCGATCTNNNNNDrC	1	NNNNNDC
c3t_ddCD5N	/5Cy5/ACACTCTTTCCCTACACG ACGCTCTTCCGATCTNNNNND/3 ddC/	0	NNNNNDC
c3t_ddCA5N	/5Cy5/ACACTCTTTCCCTACACG ACGCTCTTCCGATCTNNNNNA/3 ddC/	0	NNNNNAC
c3t_ddCG5N	/5Cy5/ACACTCTTTCCCTACACG ACGCTCTTCCGATCTNNNNNG/3 ddC/	0	NNNNNGC
* c3t_ddCR5N	/5Cy5/ACACTCTTTCCCTACACG ACGCTCTTCCGATCTNNNNNR/3 ddC/	0	NNNNNRC
**c3t_ddCR5N_barcoded	/5Cy5/ACACTCTTTCCCTACACG ACGCTCTTCCGATCTNNNNNXX XXXXNNNNNR/3ddC/	0	NNNNNXXX XXXNNNNN RC

Notes: Names, sequences, and descriptions of adapter oligonucleotide sequences used in this study. Citations where these were previously introduced included below the names where applicable.

*:Current recommendation for OTTR and what is described as in Buffer 4B2.

**:Current recommendation for barcoded OTTR 3′ddC adapter template where a constant sequence to define, *e.g.,* a specific well- or cell-identifier were defined by “X.”

Table 2.6: Alternative primer duplexes to be used for capture of RNAs with specific 3′ ends.

Duplex	Name	Sequence	RNA strand	Mix ratio	Notes
+1 G	+G.v3	/5IRD800/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTG	RNA.v3 or RNA.v3.1	1:1	Excessive levels of adapter-dimer cDNA will accumulate.
+1 A	+A.v3	/5IRD800/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTA	RNA.v3 or RNA.v3.1	1:1	
+1 T	+T.v3	/5IRD800/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTT	RNA.v3 or RNA.v3.1	1:1	
+1 C	+C.v3	/5IRD800/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTC	RNA.v3 or RNA.v3.1	1:1	
Blunt	+T.v2	/5IRD800/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	RNA.v3 or RNA.v3.1	1:1	dG and dA NTA to blunt duplexes are most common is current dNTP mix.
RNA passenger	RNA.v3	rArGrArUrCrGrGrArArGrAmGmCmAmCmAmCmGmUmCmUmGmAmAmCmUmCmCmAmGmU/3SpC3/	2′-O-methyl ribonucleotides are immune to RNase H, can potentially interfere with cDNA migration during electrogelphoresis.		
	RNA.v3.1	rArGrArUrCrGrGrArArGrAmGmCmAmCmAmCmGmUrCrUrGrArArCrUmCmCmAmGmUmCmAmC/3SpC3/	RNA ribonucleotides introduced between 2′-O-methyl ribonucleotides to ensure sensitivity to RNase H.		

Notes: The primer duplex 3′ overhanging for DNA oligonucleotides was included in the first column, except for the RNA passenger oligonucleotides. The second and third columns indicated the name of the oligonucleotide and sequence. For the DNA oligonucleotides, the two options for the RNA passenger strand (column four) and oligonucleotide combination ratio for DNA:RNA annealing (column five) were defined. Notes about specific +1 primer duplex or RNA oligonucleotide were included as the final column.

Chapter 3: Streamlined and sensitive mono- and di-ribosome profiling in yeast and human cells

Contents of this chapter were originally published on *Nature Methods* on October 2nd, 2023 as:

Ferguson L, Upton HE, Pimentel SC, Mok A, Lareau LF, Collins K, Ingolia NT. Streamlined and sensitive mono- and di-ribosome profiling in yeast and human cells. *Nat Methods*. 2023. doi: 10.1038/s41592-023-02028-1. PMID: 37783882.

Abstract

Ribosome profiling has unveiled diverse regulation and perturbations of translation through a transcriptome-wide survey of ribosome occupancy, read out by sequencing of ribosome-protected mRNA fragments. Generation of ribosome footprints and their conversion into sequencing libraries is technically demanding and sensitive to biases that distort the representation of physiological ribosome occupancy. We address these challenges by producing ribosome footprints with P1 nuclease rather than RNase I and replacing RNA ligation with Ordered Two-Template Relay, a single-tube protocol for sequencing library preparation that incorporates adapters by reverse transcription. Our streamlined approach reduced sequence bias and enhanced enrichment of ribosome footprints relative to ribosomal RNA. Furthermore, P1 nuclease preserved distinct juxtaposed ribosome complexes informative about yeast and human ribosome fates during translation initiation, stalling, and termination. Our optimized methods for mRNA footprint generation and capture provide a richer translatome profile with low input and fewer technical challenges.

Introduction

Ribosome profiling provides a global, high-resolution view of translation by deep sequencing of ribosome-protected mRNA fragments (RPFs). By counting the number of RPFs along each protein coding sequence (CDS), a snapshot of active cellular translation can be measured across the proteome at the point of cell lysis. Careful analysis of the position of each RPF enables deeper interpretations of ribosome occupancy, such as the decoding rate of individual codons. This approach has yielded insights about translation mechanisms, non-canonical events such as frameshifting and reinitiation, and open reading frame (ORF) discovery (Ingolia et al. 2009, Gao et al. 2015, Irigoyen et al. 2016, Napthine et al. 2017, Irigoyen et al. 2018, Stewart et al. 2018, Cook et al. 2022, Lulla et al. 2019). Further insights have been gleaned from adaptations of ribosome profiling that capture specific ribosome sub-populations, pre-initiation complexes, and disomes produced by ribosome-ribosome collisions (Oh et al. 2011, Jan et al. 2014, Williams et al. 2014, Döring et al. 2017, Archer et al. 2016,

Wagner et al. 2020, Bohlen et al. 2020, Guydosh and Green et al. 2014). Despite these benefits, widespread use of ribosome profiling has been hampered by labor-intensive workflows with high RNA input requirements and excessive capture of ribosomal RNA (rRNA) fragments generated during nuclease digestion (Ingolia et al. 2009).

Ribosome profiling has two essential steps: treating cell lysate with a nuclease that degrades accessible mRNA to generate RPFs, and converting these RPFs into a complementary DNA (cDNA) library for high-throughput sequencing. Each step poses technical challenges. The ideal nuclease would cleave unprotected mRNA with no sequence bias while sparing the structural RNA of the ribosome. The original and most widely used nuclease, *Escherichia coli* RNase I, has no nucleotide preference but readily digests rRNA, leading to substantial rRNA contamination (Ingolia et al. 2009). Careful titration of RNase I or alternative nucleases are needed for very low-input samples (Xiong et al. 2022, Li et al. 2022a, Zhang et al. 2022, VanInsberghe et al. 2021). Alternatives such as micrococcal nuclease (MNase), RNase A, and RNase T1 can reduce rRNA contamination and better preserve ribosomes, but they have strong RNA sequence preferences so they do not provide the single-nucleotide resolution of RNase I RPFs and may distort measurements of translation speed and ribosome stalling (Gerashchenko and Gladyshev. 2017). Likewise, the ideal library generation approach would convert RPFs into a sequencing library efficiently and with little bias. Existing workflows ligate the 3′ ends of RPFs to adapter oligonucleotides, an inefficient approach with sequence biases (Tunney et al. 2018). SMARTer-seq small RNA sequencing has been adapted to sequence RPFs (Hornstein et al. 2016), although framing information can be lost. Moreover, use of ligase obliges denaturing gel purifications to separate product from unreacted substrates.

We recently reported a straightforward and low-bias approach to construct small RNA sequencing libraries that takes advantage of the enzymatic activities of reverse transcriptase (RT) enzymes from eukaryotic retroelements. We used the terminal transferase and template jumping cDNA synthesis activities of an engineered retroelement RT to develop an Ordered Two-Template Relay (OTTR) protocol for small RNA library generation (Upton et al. 2021). This rapid and single-tube workflow eliminates any need for RNA or DNA ligases, replacing them with a programmed sequence of template jumps that directs continuous cDNA synthesis from a 5′ adapter duplex primer across the input RNA template and then a 3′ adapter template (Upton et al. 2021). On microRNA templates, OTTR displays substantially less sequence bias and more even coverage than SMARTer-seq (Upton et al. 2021), and similar benefits are expected for RPFs.

Here we describe a comprehensive redevelopment of the widely used ligation-based ribosome profiling workflow with reduced effort, less complexity, and low input requirements. We first established OTTR as a suitable replacement for RPF sequencing library generation by direct comparison with the ligation-based approach (McGlincy and Ingolia 2017). We then developed and validated P1 nuclease — a sequence-independent, single-strand nuclease — as an alternative to RNase I for RPF production in yeast and human cell lysates. We found that P1 nuclease preserved ribosome integrity better than RNase I and reduced rRNA contamination in RPF libraries. P1

nuclease footprinting and OTTR library synthesis detected the ribosomal stalling induced by impaired translation, and further revealed several configurations of adjacent, collided ribosomes, which trigger ribosome and/or mRNA turnover by ribosome quality control (RQC) pathways. Lastly, we further streamlined the combined P1 nuclease and OTTR ribosome profiling workflow by eliminating gel electrophoresis during RPF recovery, instead using a small RNA enrichment approach, mirRICH. Taken together the workflow simplifications and improved information gain will enable broader adoption of ribosome profiling.

Results

High-yield and low-bias ribosome profiling with OTTR

OTTR provides a streamlined, efficient, and low-bias small RNA library generation strategy that might both simplify and improve on current methods used in ribosome profiling. In place of sequential ligations, OTTR proceeds through terminal transfer of a dideoxynucleotide onto the small RNA template followed by template jumping cDNA synthesis (**Fig. 3.1A**) (Upton et al. 2021). A defined 5′-primer is extended first using a small RNA template followed by a constant 3′ adapter template; distinct 1-nt overhangs enforce the order of template jumps to drive the synthesis of a cDNA product with adapter sequences on each side of a small RNA insert, suitable for high-throughput sequencing.

To assess whether OTTR library generation is beneficial for ribosome profiling, we compared it to a standard ligation-based workflow (McGlincy and Ingolia 2017). We prepared RPFs from *Saccharomyces cerevisiae* (budding yeast) by RNase I digestion and split them into two portions to compare the two workflows; because OTTR can generate libraries from low input (Upton et al. 2021), we used ten-fold less RPFs for OTTR library generation than for the ligation-based workflow. We generated OTTR ribosome profiling libraries in a single-tube OTTR reaction over the course of about 4 hours, and created ligation-based libraries by following the typical 3-day workflow (**Fig. 3.1A**). The ligation-based method converted ~400 ng of input RNA to 24 fmoles of DNA library while OTTR converted ~40 ng to 180 fmoles of library. Further, the OTTR library contained a lower fraction of unwanted library generation byproducts such as adapter-only reads (**Extended Data 3.1A**). The gene- and codon-level RPF occupancies were highly correlated, and we concluded that OTTR provided comparable data to the ligation-based workflow (**Fig. 3.1B-C**).

A strength of ribosome profiling is the ability to infer the codon being decoded by each ribosome based on the length and position of each RPF. The peptidyl tRNA (P) and acceptor tRNA (A) sites are typically offset 12 – 14 and 15 – 17 nucleotides (nt), respectively, from the 5′ end of an RNase I RPF. This information, coupled with the 3-nt periodicity of ribosome elongation, allows us to determine the A-site position of each RPF (Ingolia et al. 2009). These A-site positions can be calibrated empirically from ribosome occupancy at initiation and termination codons (**Fig. 3.1D**), but library

generation artifacts may impact this assignment. We noted that RPF sequences from the OTTR library were slightly shorter than those from the ligation-based library, despite originating from the same RPF pool, due to two technical effects (**Extended Data 3.1B**). The ligation-based library included 29 nt reads with high mismatch rates at the 5′-most position, a known consequence of non-templated nucleotide addition by retroviral RTs (**Extended Data 3.1C**) (McGlincy and Ingolia 2017). These extended reads were essentially absent in the OTTR library. Second, OTTR adds an extra adenosine dideoxynucleotide to the 3′ end of an RPF, which is then computationally trimmed. Some RPFs do not receive this extra nucleotide and are nonetheless captured into the cDNA library (**Extended Data 3.1D-E**), so computational trimming artifactually shortens these reads; we are currently optimizing terminal transferase conditions to reduce this artifact. Despite these minor differences, both methods captured the characteristic triplet periodicity of RPFs indicative of translation reading frames (**Fig. 3.1D-E)**.

Technical factors such as ligase sequence preferences in ribosome profiling library generation have been shown to distort measurements of ribosome occupancy (Tunney et al. 2018). For example, the ligation-based library, but not the OTTR library, had a translatome-wide peak of read density for footprints with 5′ ends at the start codon, corresponding to ribosomes on the 5th codon (**Fig. 3.1D-E**). We and others have noted this enrichment in many ribosome profiling data sets using ligation for cDNA library production (Ingolia et al. 2009, McGlincy and Ingolia 2017, Han et al. 2014). The absence of this 5th-codon peak of read density in the OTTR library suggested that it could arise from a technical bias rather than an accumulation of ribosomes at this position in cells. Indeed, because all of these 5th codon RPFs have the same nucleotide sequence at their 5′ end, namely the AUG start codon, the sequence bias of enzymes used in the ligation-based protocol could explain their over-representation.

To quantify factors that influence RPF abundance in OTTR versus ligation-based libraries, we evaluated each library using iχnos, a neural network model of ribosome profiling data that predicts ribosome distribution along a transcript based on the sequences in a defined window encompassing the entire RPF (Tunney et al. 2018). Along with predicting the ribosome occupancy at each position, iχnos evaluates the predictive power of sequence features around that position. The identities of the codons in the A and P sites, which are paired with tRNAs within the ribosome, are expected to have the biggest impact on the speed of translation and thus the accumulation of RPFs. However, previous analysis with iχnos found that, in many RPF libraries made by ligation, the density of RPFs at a position was driven largely by the sequences at the 5′ and 3′ ends of the fragments, reflecting a strong influence of enzyme bias in RPF capture rather than the biology of ribosome decoding (Tunney et al. 2018). In keeping with this, models trained on our ligation-based RPF libraries showed the strongest effect from the identity of the sequences at the 5′ end of the RPF, confirming the strong bias introduced by CircLigase II (**Fig. 3.1F**). In contrast, we found the A-site and P-site codons contributed most to RPF abundance in OTTR libraries (**Fig. 3.1F**). We conclude that the abundance of individual RPFs captured by OTTR was more indicative of tRNA-paired codon identities, and by inference, ribosome dwell time while decoding specific codons.

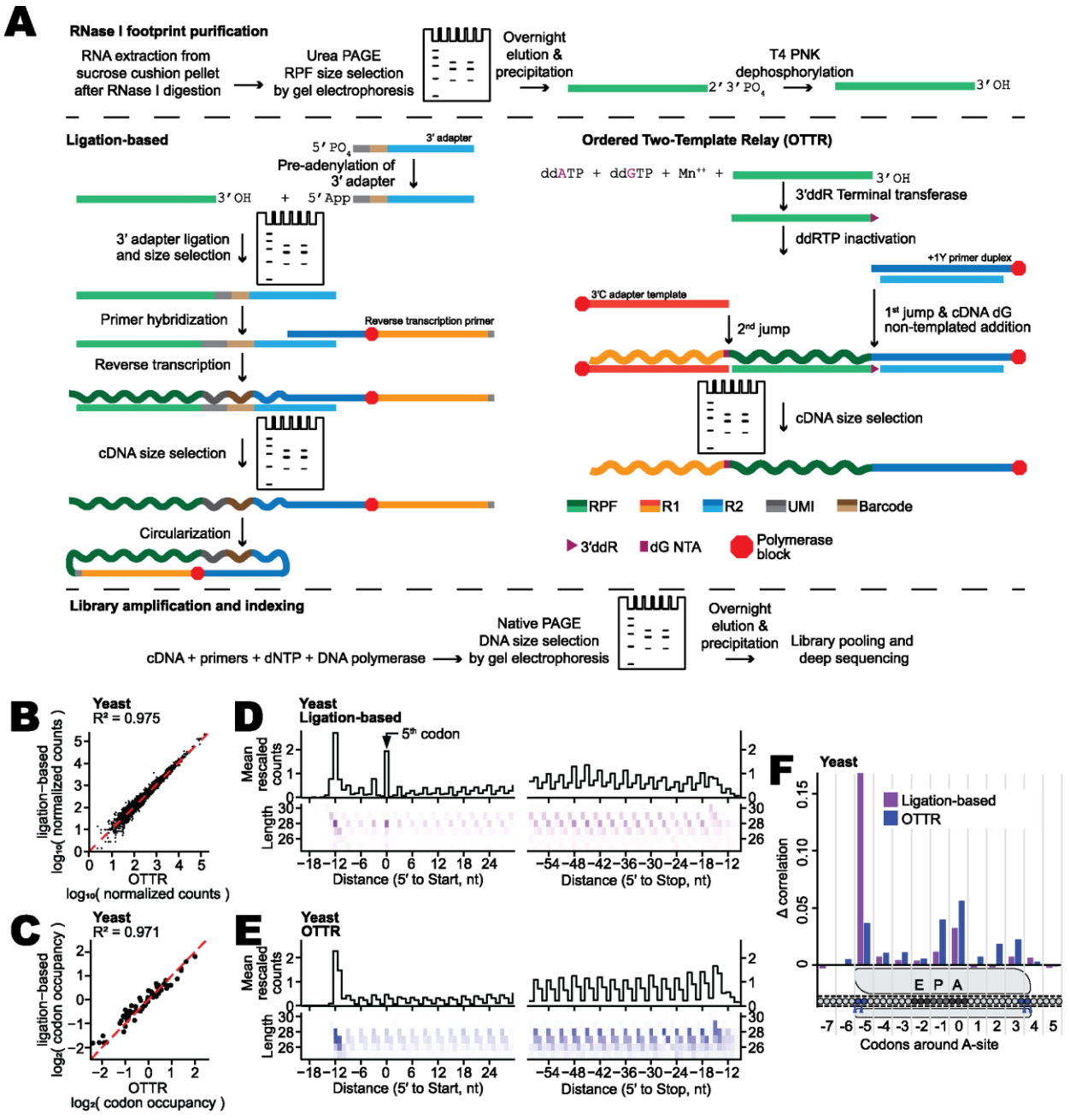


Figure 3.1: Highly correlated ribosome profiles derived from ligation-based or OTTR library generation.

A, Schematic of library generation by OTTR or ligation-based protocols from a single pool of RNase I derived RPFs (green) from a sucrose cushion. The pool of RPFs were split unevenly after T4 polynucleotide kinase (PNK) treatment with only 1:10th of the RPFs used in OTTR. In OTTR, each step before cDNA size selection occurs within 4 hours in a single tube. First, input RPF RNA was labeled by either ddA or ddG on the 3′ end before unincorporated ddRTPs were inactivated by rSAP. Then, two ordered jumps,

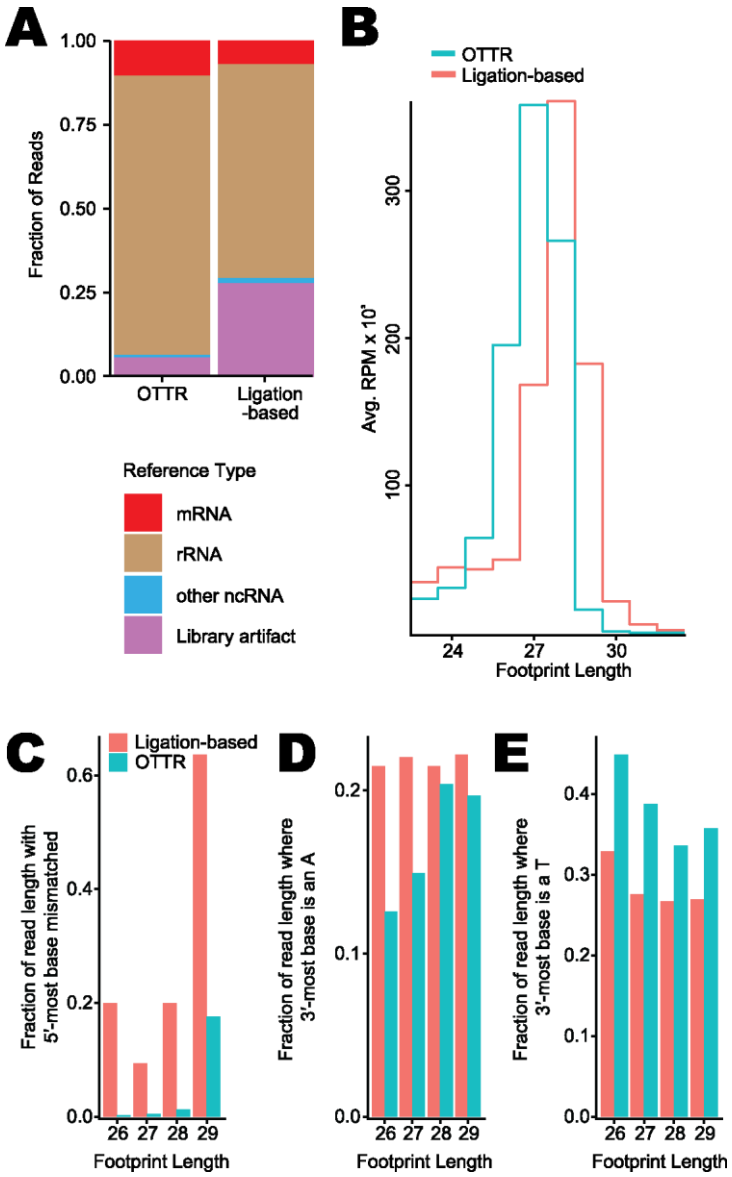
the first initiated from the +1Y DNA/RNA primer duplex to the input RPF RNA, and the second initiated from a non-templated dG addition to the RPF cDNA to jump to the 3′C adapter template, yields a cDNA molecule with a 5′ and 3′ adapter flanking the complement of the RPF input. In ligation-based, the 3′ adapter is first adenylated on its 5′ end. Then, 3′ adapter ligation is carried out, followed by gel-based size selection and overnight elution. The next day, material is precipitated and primer hybridization for reverse transcription occurs. Following reverse transcription, cDNA is purified by gel-base size selection. After elution, cDNA is circularized. In these illustrations green/light green denoted the RPF sequence, orange/light orange denoted the R1 adapter sequence, blue/light blue denoted the R2 adapter sequence, gray/dark gray denoted the unique molecular identifier sequence, brown/light brown denoted the barcode sequence, red octagon denoted polymerase blocking groups, magenta triangle denoted a 3′ddR, and a magenta square denoted the dG non-templated addition.

B, Comparison of gene-level ribosome occupancy estimates from libraries generated in (**A**). Read counts are for RPFs aligned to verified CDSs excluding those RPFs that are aligned to the first 15 and last 10 codons. Read counts for each gene were normalized by DESeq2. Two pools of T4 PNK-treated RPFs were prepared by ligation-based on OTTR, but only one of the two libraries was considered successful from a ligation-based workflow. Pearson's R^2 is reported for all codons

C, Comparison of mean codon-level occupancy estimates from libraries generated in (**A**). Aligned RPFs were assigned to an A-site codon and counted. These counts were then rescaled by the mean codon count for the gene, excluding those RPFs that are aligned to the first 15 and last 10 codons, and averaged across the translatome. Pearson's R^2 is reported for all codons

D-E, Metagene profile around the start (left) and stop (right) codons for either (**D**) OTTR or (**E**) ligation-based libraries. Aligned RPFs for each CDS were first rescaled by the mean codon count for the gene, excluding RPFs aligned to the first 15 and last 10 codons, and then averaged across the translatome. Footprints were tabulated according to either the 5′ aligned position alone (shown at top as a black line), or both 5′ aligned position and read length (shown at bottom as a matrix of distinct RPF lengths and positions).

F, Per-codon contributions to iχnos machine learning models of RPF occupancy profiles. A model based on a widow of 13 codons (-7 to +5) around the A-site was compared with thirteen additional models, each omitting one codon from the model. The contribution of a codon position to RPF occupancy profile was inferred from the change in Pearson's correlation coefficient between the predicted ribosome occupancy versus actual ribosome occupancy changed when the codon was omitted (Y-axis).



Extended Data 3.1: Additional comparison of RNase I RPF libraries by ligation-based or OTTR protocol.

A, Fraction of RNase I RPF sequencing library reads that mapped to each transcript class. Library generation artifacts included sequences that were adapter-only, shorter than 15 bases, or unmapped.

B, CDS RPF length distribution from OTTR (blue) and ligation-based (red) RNase I ribosome profiling libraries. RPFs aligned to the first 15 or final 10 codons were excluded. Counts were represented in reads per million mapped reads (RPM) and averaged across replicates

C, For each read length from 26 to 29 nt, the fraction of RPF alignments with

mismatches at the 5′-most base of the alignment. For this analysis, alignments were permitted to only have a single mismatch to the reference.

D, For each read length from 26 to 29 nt, the fraction of RPF alignments with an adenosine (A) at the 3′-most base of the alignment. For this analysis, alignments were permitted to only have a single mismatch to the reference.

E, For each read length from 26 to 29 nt, the fraction of RPF alignments with a thymine (T) at the 3′-most base of the alignment. For this analysis, alignments were permitted to only have a single mismatch to the reference.

P1 nuclease offers advantages over RNase I for footprinting

E. coli RNase I has advantages in ribosome footprint generation, relative to alternatives such as combinations of RNases A and T1 or MNase, primarily due to a lack of sequence preference (Gerashchenko and Gladyshev. 2017). However, RNase I digestion extensively fragments the rRNA, leading to two issues: 1) contamination of sequencing libraries with RPF-sized rRNA fragments; and 2) compromised structural integrity in a fraction of ribosomes, leading to subunit dissociation. Indeed, careful RNase I titration is a key optimization in recent low-input ribosome profiling workflows (Xiong et al. 2022, Li et al. 2022a, Zhang et al. 2022). We sought a sequence-independent RNase that would preferentially cleave unprotected mRNA while better sparing highly structured rRNA. The S1/P1 family of endonucleases stood out for its lack of nucleotide specificity combined with strong preference for single-stranded RNA and production of cleavage products with 3′-terminal hydroxyls (Koval’ et al. 2016, Volbeda et al. 1991). Although they are generally active at acidic pH, here we report P1 nuclease from *Penicillium citrinum* retains activity in a pH range compatible with ribosome stability in cell extracts (Fujimoto et al. 1974).

To determine P1 nuclease digestion conditions suitable for ribosome profiling, we tracked the extent of unprotected mRNA cleavage in crude cell lysates by measuring the conversion of polysomes to monosomes (**Extended Data 3.2A-B**). P1 nuclease digestion optimally separated yeast polysomes into monosomes when we lysed cells in standard pH 7.5 buffer then adjusted the pH of the lysate to 6.5 before adding P1 nuclease and incubated at 30 °C for one hour (**Extended Data 3.2A**). Compared to RNase I digestion of an equivalent input sample, P1 nuclease digestion preserves more 80S monosomes and a small population of nuclease-resistant disomes (**Fig. 3.2A**). P1 nuclease digestion did not produce free 40S and 60S subunits, whereas RNase I did generate free subunits, raising the possibility that it cleaved rRNA to the point of causing subunit dissociation (**Fig. 3.2A**). We likewise found that P1 nuclease digestion of lysates from cultured human Calu-3 cells converted polysomes into monosomes with a small nuclease-resistant disome peak and minimal changes to free 40S and 60S subunits (**Fig. 3.2B, Extended Data 3.2C**). Both yeast and human cell lysate showed an optimal plateau of monsome yield centered on 15 U / µg RNA (**Fig. 3.2C-D**), with digestion

performed at their respective growth temperatures of 30 ºC and 37 ºC (**Extended Data 3.2B-C**).

We next assessed rRNA integrity following RNase I or P1 nuclease digestion. We anticipated that limiting rRNA cleavage would reduce RNA within the RPF size range. After digestion, we recovered ribosomes by ultracentrifugation through a sucrose cushion, extracted RNA, and characterized it by denaturing PAGE. This revealed a stark contrast in the RNA fragment profiles between RNase I and P1 nuclease digestion: RNase I generates a range of abundant small RNA fragments in the range of RPF lengths, whereas P1 nuclease produces far less RNA in this size range (**Fig. 3.2E**). This observation, coupled with the limited creation of free 40S and 60S subunits, suggests that P1 nuclease preferentially cleaves mRNA over rRNA relative to RNase I.

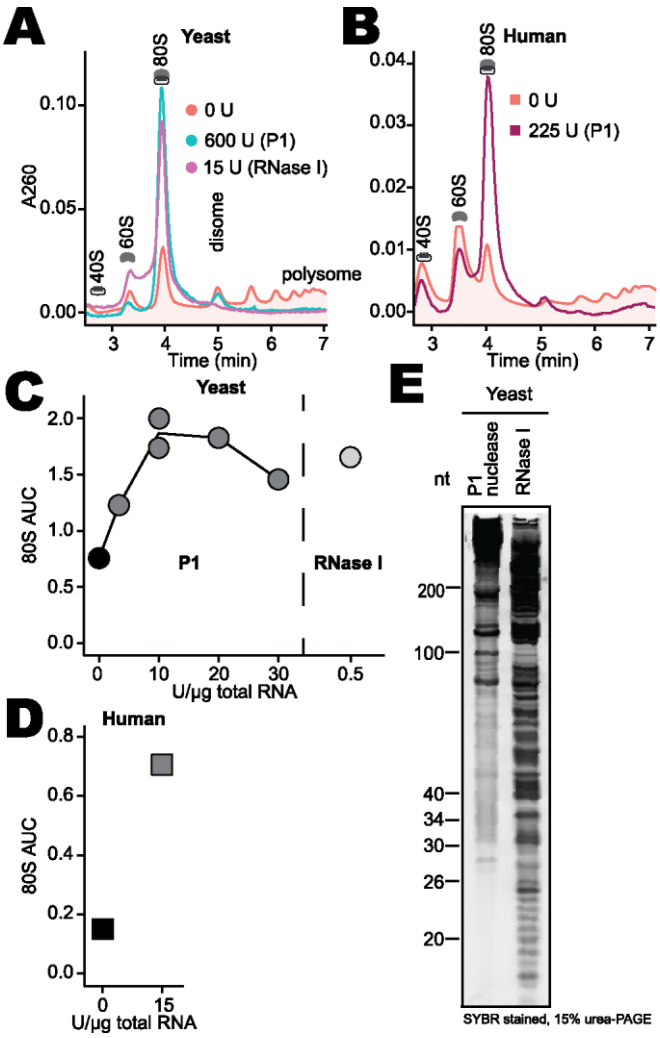


Figure 3.2: P1 nuclease collapse of polysomes into monosomes while limiting rRNA degradation.

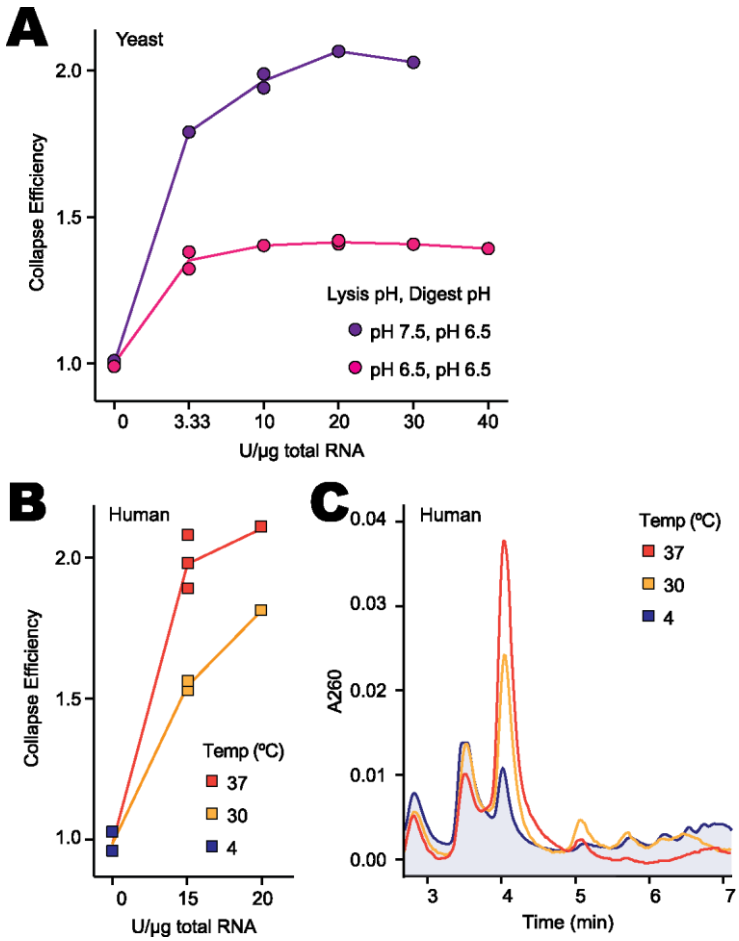
A, Sucrose density gradient polysome profiles of nuclease-treated yeast lysate, along with an undigested control. Each sample contained 30 µg of total RNA in 200 µL prior to nuclease digestion with the units (U) of enzyme indicated. P1 nuclease samples were adjusted to pH 6.5 prior to digestion and digested at 30°C for 1 hour. RNase I digestion was at room temperature for 45 min.

B, As in (**A**), for human-derived Calu-3 cell lysate containing 15 µg of total RNA.

C, Integrated monosome peak area from yeast lysate digested with various concentrations of P1 nuclease at 30 °C or RNase I at room temperature for an hour. Dark fill indicates undigested control, which was incubated without nuclease at 4 °C for an hour (n=1 for each condition, except n=2 for 10 U/µg).

D, As in (**C**) for human Calu-3 cell lysate, but with P1 nuclease digestion at 37 °C. Dark fill indicates undigested control, which was incubated without nuclease at 4 °C for an hour (n=1 for each condition).

E, Denaturing PAGE analysis of RNA extracted from sucrose cushioned pellets after digestion with P1 nuclease or RNase I, as in (a), and stained with SYBR Gold. One replicate of three for both conditions was shown.



Extended Data 3.2: Optimization of P1 nuclease digestion and cell lysis conditions.

A, Polysome collapse efficiency analysis comparing lysis with either pH 6.5 or pH 7.5 polysome lysis buffer and addition of a range of P1 nuclease U/µg. All assays were carried out with lysate measuring 30 µg of total RNA diluted to 200 µL with either pH 6.5 or pH 7.5 polysome buffer. At the point of nuclease digestion, pH 7.5 lysate was adjusted to ~pH 6.5 with 14 µL of 300 mM Bis-Tris (pH 6.0), and pH 6.5 lysate was supplemented with an additional 14 µL of pH 6.5 polysome buffer. Collapse efficiency was calculated as the ratio of the integrated monosome peak absorbance relative to the integrated polysome region absorbance, normalized by the undigested control (n=1 for

each condition, except n=2 for pH 6.5 with 3.33 U/µg or 20 U/µg and pH 7.5 with 0 U/µg or 10 U/µg).

B, Comparison of polysome collapse efficiency by P1 nuclease digestion at either 30 °C or 37 °C using Calu-3 human cell lysate. Briefly, lysate measuring 15 µg of total RNA was diluted to 200 µL in pH 7.5 polysome buffer and pH adjusted with 14 µL of 300 mM Bis-Tris (pH 6.0) before supplemented with P1 nuclease and digestion at 30 °C or 37 °C. Undigested control incubated at 4 °C for an hour without nuclease (n=2 for 4 ºC no-nuclease controls, n=2 or 3 for 15 U/µg as shown, and n=1 for 20 U/µg).

C, Representative polysome profile from P1 nuclease digestion of Calu-3 human cell lysate at 30 ºC or 37 °C with nuclease at 15 U/µg total RNA. Undigested control incubated at 4 °C for an hour without nuclease.

Ribosome profiling with P1 nuclease footprinting

To directly compare the RPFs produced by RNase I and P1 nuclease digestions, we generated pools of RNase I or P1 nuclease RPFs from the same yeast and human 293T cell lysates. We generated cDNA libraries using OTTR (**Extended Data 3.3A**) and sequenced them. P1 RPFs were longer than RNase I RPFs: P1 RPFs were 31 – 36 nt in yeast, compared to 26 – 29 nt RNase I RPFs, and in human cells P1 RPFs were 34 – 39 nt, compared to 28 – 32 nt for RNase I RPFs (**Extended Data 3.3B-C**). P1 RPF libraries yielded a larger fraction of reads mapping to mRNA (**Extended Data 3.3D-E**) and contained fewer discrete rRNA fragment sequences of high abundance (**Extended Data 3.3F-H**). Of note, P1 nuclease produced more tRNA-derived reads relative to RNase I, particularly in human cells (**Extended Data 3.3D-E**).

In OTTR libraries made with RNase I or P1 nuclease, we saw high ribosome occupancy around translation initiation sites followed by footprint coverage across the CDSs with clear triplet periodicity (**Fig. 3.3A-B**). Gene-level ribosome occupancy measurements showed high correlation between nucleases for both yeast and human libraries (**Fig. 3.3C-D**). P1 RPFs from initiating yeast and human ribosomes had 5′ ends extended by 1 – 2 nt compared to RNase I footprints, positioning the A-site codon 16 – 17 nt from the 5′ end (**Fig. 3.3E-F**). This increase in RPF length using P1 nuclease versus RNase I was also evident in RPFs from terminating yeast ribosomes: RNase I produced RPFs with 5′ ends located 14 – 15 nt upstream of the stop codon while P1 RPFs had 5′ ends 16 – 17 nt upstream (**Fig. 3.3A**). The heterogeneity of RPF lengths for any given ribosome A-site position primarily reflects a heterogeneous position of RPF 3′ end (**Fig. 3.3E-F**).

We used the RPFs at initiating codons to empirically calibrate the offset between the 5′ ends of a footprint and the A- and P-site codons for both RNase I and P1 nuclease footprints. Across CDSs, inferred A-site codon occupancies correlated well for yeast RPFs produced by P1 nuclease and RNase I (R^2 = 0.884) (**Fig. 3.3G**). The correlation was somewhat lower between P1 nuclease and RNase I from human cell

lysates (R^2 = 0.471) driven primarily by an under-representation of arginine A-site codon assignments among RNase I RPFs (**Fig. 3.3H**). We selected RNase I RPFs of 26 – 34 nt, which would exclude the ~21 nt RPF that arise from RNase I cleavage within an A-site when unoccupied by a tRNA (Lareau et al. 2014, Wu et al. 2019). Codon-level differences between RNase I and P1 RPFs could reflect differences in empty A-site cleavage: RNase I cleaves within the A-site at slowly-decoded arginine codons and truncates the RPF, while P1 nuclease leaves these RPFs intact.

Indeed, several lines of evidence suggest that P1 nuclease is more sensitive to steric hindrance and thus less likely to cleave within the A-site. P1/S1 nuclease makes extensive contacts with several single-stranded nucleotides upstream of the cleaved phosphodiester bond (Koval' et al. 2016, Romier et al. 1998). Moreover, P1 nuclease cleavage of DNA mismatches requires 2 – 3 nt of consecutive single-stranded nucleotides (Naik and Raghavan 2008). We propose that P1 nuclease requires several nucleotides of unobstructed upstream mRNA in order to cleave. This asymmetry explains the uniform cleavage on the 5′ side of the footprint and the wider range of 3′ cleavage positions, and thus, RPF lengths (**Fig. 3.3I**).

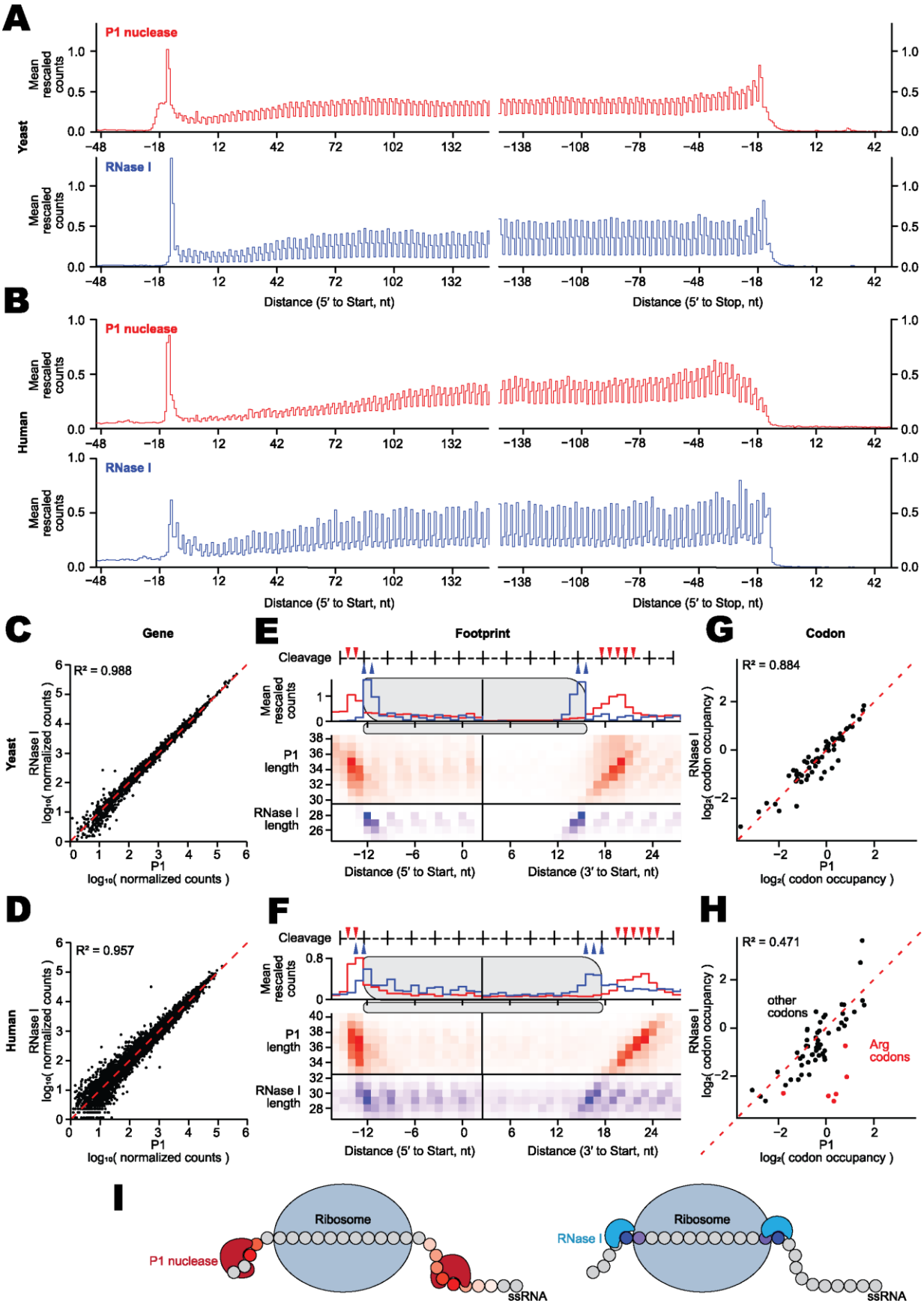

A
P1 nuclease
RNase I
Yeast
Mean rescaled counts
Distance (5′ to Start, nt)
Distance (5′ to Stop, nt)
B
Human
C
Gene
R² = 0.988
Yeast
RNase I log₁₀(normalized counts)
P1 log₁₀(normalized counts)
D
R² = 0.957
Human
E
Footprint
Cleavage
P1 length
RNase I length
Distance (5′ to Start, nt)
Distance (3′ to Start, nt)
F
G
Codon
R² = 0.884
RNase I log₂(codon occupancy)
P1 log₂(codon occupancy)
H
R² = 0.471
other codons
Arg codons
I
Ribosome
P1 nuclease
RNase I
ssRNA

Figure 3.3: Highly similar ribosome profiling data from P1 nuclease or RNase I digestion.

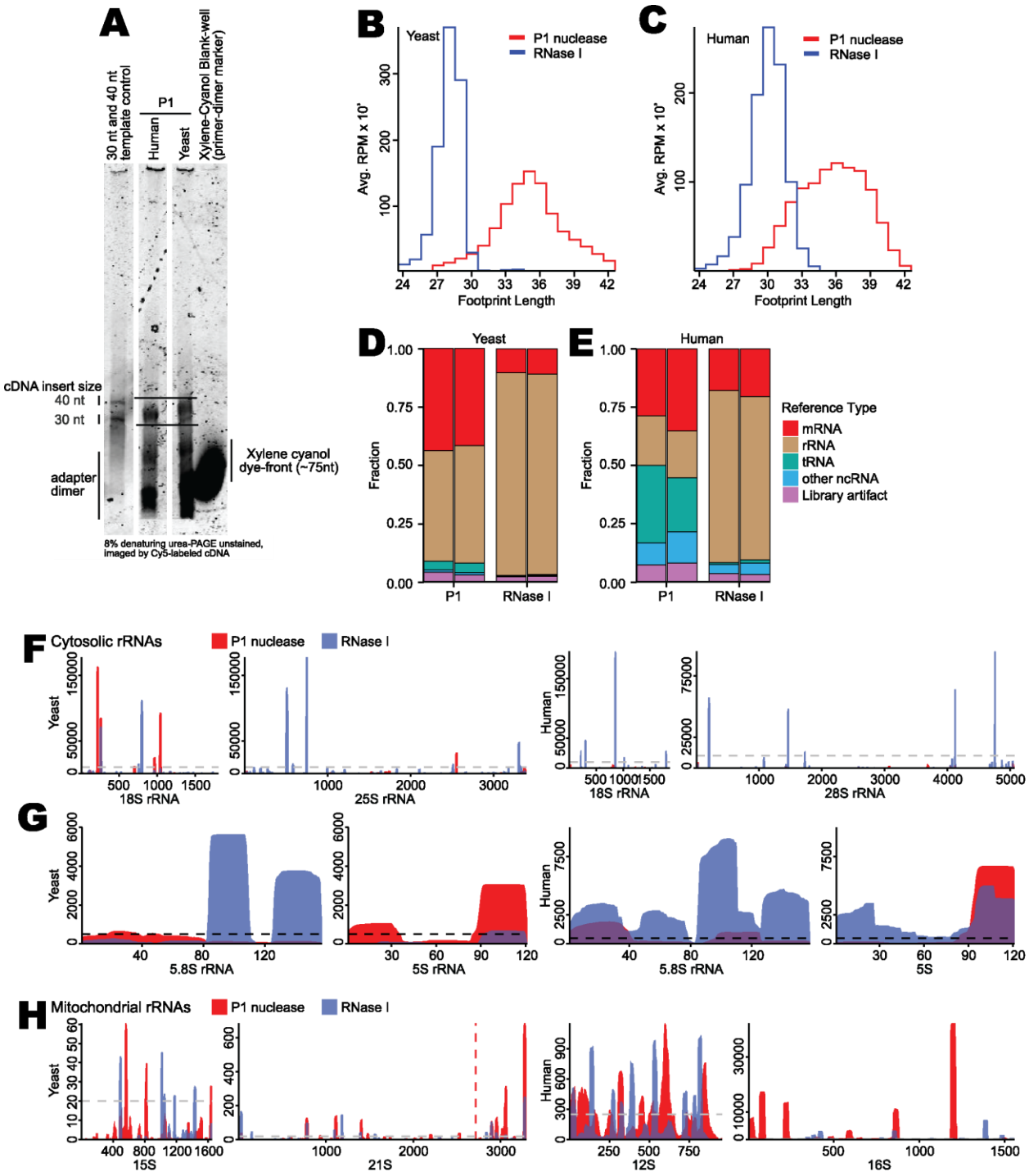
A-B, Metagene average profiles around the start (left) and stop (right) codons from sucrose cushion purified yeast (**A**) and human 293T cell RPFs (**B**) generated by P1 nuclease (top, red) or RNase I (bottom, blue) digestion, as in **Fig. 1D-E** (top). Two libraries were made for each condition; data was combined for metagene profiles.

C-D, Gene-level ribosome occupancy estimates normalized by DESeq2, as in **Fig. 1B**, from sucrose cushion purified yeast RPFs (**c**) or human 293T cell RPFs (**D**) generated by P1 nuclease and RNase I digestion (n=2 libraries for each condition).

E-F, Mean yeast (**E**) and human 293T cell (**F)** ribosome occupancy at start codons for P1 nuclease (red, 30 – 40 nt for yeast; or 33 – 40 nt for human) and RNase I (blue, 25 – 29 nt for yeast; or 27 – 32 nt for human) libraries. Read counts were rescaled separately for both 5′ end (left) and 3′ end positions (right), prior to averaging across all genes, as in (**A**). Heatmap of the mean rescaled counts according to length and end position (below), and stepped line plot for the summed end position (above each heatmap matrix). The 5′ and 3′ end averages are shown to the left and to the right, respectively, of a black vertical bar. A diagram of a translating ribosome footprint (top) indicates mRNA cleavage positions of P1 nuclease (red triangle) and RNase I (blue triangle).

G-H, Comparison of codon-level ribosome occupancy estimates, as in **Fig. 3.1C**, from (**G**) yeast or (**H**) human 293T cell RPFs generated by P1 nuclease and RNase I digestion. In (**H**), arginine codons are shown in red.

I, Schematic of proposed P1 nuclease and RNase I cleavage sites around an mRNA-engaged ribosome. Increased frequency of an RPF terminal position is indicated by increasing color saturation.



Extended Data 3.3: OTTR library production and ribosome profiles.

A, A 0.6X TBE 8% denaturing urea-PAGE gel of OTTR library cDNA generated from human 293T and yeast cell P1 nuclease RPFs purified by sucrose cushion. The cDNA was directly imaged by the Cy5 dye covalently linked to the ORRT adapter duplex (see **Fig. 3.1A**). A parallel OTTR library was synthesized from size marker oligonucleotides

(**Table 3.1**). Xylene cyanol co-migrates with OTTR adapter dimer cDNA. Horizontal lines indicate excision.

B-C, CDS RPF length distribution from yeast (**B**) and human 293T (**C**) RNase I (blue) and P1 nuclease (red) ribosome profiling libraries, as in **Extended Data 3.1B**.

D-E, Fraction of sequencing reads mapped to each transcript class yeast (**D**) and human 293T (**E**) ribosome profiling libraries, as in **Extended Data 3.1A**.

F, Mean per-base read coverage of cytosolic 18S and 25S (or 28S) rRNA from yeast (left) or human (right) ribosome profiles made from P1 nuclease (red) or RNase I (blue) digestion. Coverage was represented in reads per million total mapped reads (i.e., rRNA, tRNA, ncRNA, mRNA, and other genomic loci)

G, As in (**F**) but for 5.8S and 5S rRNA coverage.

H, As in (**F**) for mitochondrial rRNA coverage.

P1 nuclease creates disome footprints from apposed ribosomes

In light of this model for P1 nuclease cleavage, we wanted to know whether P1 could cleave between closely spaced ribosomes. Intriguingly, we noted the presence of a nuclease-resistant disome peak in the polysome profiles yeast and human cell lysates after P1 digestion (**Fig. 3.2A-B**). Adjacent ribosomes are known to arise when one ribosome stalls during elongation and a second, trailing ribosome collides with it. Such ribosome-ribosome collisions trigger quality control and reduce overall translation initiation through the integrated stress response (Wu et al. 2020, Yan and Zaher 2021). To test whether P1 digestion preserves long RPFs spanning two collided ribosomes, we induced translational stalling at histidine codons by depriving yeast cells of charged histidyl-tRNA (Muller et al. 2020). We knocked down histidyl-tRNA synthetase 1 (*HTS1*) mRNA in yeast that lacked Gcn2 and thus could not reduce translation initiation in response to uncharged tRNA accumulation (**Fig. 3.4A**) (Muller et al. 2020, Sonenberg and Hinnebusch 2009). Knockdown of *HTS1* increased the residual disome population following P1 nuclease digestion (**Fig. 3.4B, Extended Data 3.4A**).

Given that disomes are a minor constituent of the P1 nuclease digestion products, it was necessary to separate them from the abundant monosomes by sucrose density gradient fractionation prior to library generation (**Fig. 3.4B**). From the monosome peak, we selected 30 – 40 nt RNAs (**Fig. 3.4C, Extended Data 3.4B**). In the disome fraction, negative-stain electron microscopy confirmed ribosome pairs and, potentially, solitary ribosomes (**Fig. 3.4D**). Therefore, from the disome fraction we collected 30 – 40 nt RNA fragments as a “heavy monosome” library and 45 – 80 nt RNA fragments as a “disome” library (**Fig. 3.4C, Extended Data 3.4B**). Sequencing of this heavy monosome library revealed RPFs similar to those in the conventional “light monosome” footprint library, but strongly enriched for transmembrane protein mRNAs (**Fig. 3.4E**) (Shurtleff et al. 2018). We speculate that single ribosomes translating

transmembrane proteins may sediment more rapidly than other translating ribosomes due to their associations with protein translocation machinery or aggregation of nascent transmembrane regions (Rapoport 2007).

Reads from the disome library fell into two size classes (**Fig. 3.4C**). The larger fragments were roughly 65 nt, matching the footprint size anticipated for a pair of closely adjacent ribosomes. The smaller fragments were ~50 nt, shorter than the ~58 nt disome footprints produced by RNase I digestion (Meydan and Guydosh 2020) and thus probably not derived from two adjacent, intact 80S ribosomes. We computationally separated these RPFs into “true disome” footprints of at least 60 nt and “sub-disome” footprints shorter than 60 nt (**Fig. 3.4C**).

Hts1 depletion induced dramatic ribosome stalling on histidine codons, evidenced by a >5-fold increase in ribosome occupancy (**Fig. 3.4F**). The distribution of RPFs at isolated histidine codons (i.e., no other histidine codon in a window from -30 to +10 codons; n=19,814) showed a clear accumulation of footprint 5′ ends 16 – 17 nt upstream of the histidine codon with a length-dependent distribution of 3′ ends 14 – 21 nt downstream (**Extended Data 3.4C-D**), matching our inferred A-site position based on initiating ribosomes (**Extended Data 3.4E-F**).

True disome RPFs also accumulated at histidine codons, with 5′ ends 46 – 47 nt upstream of the histidine codon and a distribution of 3′ ends similar to monosome RPFs (**Fig. 3.4G**). We also detected a small sub-population of true disome RPFs extended by 3 nt on their 5′ ends, suggestive of a leading ribosome stalled on a histidine codon with a trailing ribosome in some cases 11 rather than 10 codons behind it (**Fig. 3.4G,** see asterisk). To investigate the P1 nuclease footprinting of nearly collided ribosomes, we examined positions where two histidine codons are separated by 10 – 13 non-histidine codons. Histidine pairs separated by 10 codons yield the same ~65 nt disome footprint that accumulates at isolated histidine codons (**Fig. 3.4H,** top left), consistent with a 10-codon separation between collided ribosomes. For longer codon gaps we observed a steady increase in footprint lengths, principally from stepwise 3 nt extension of the 5′ ends (**Fig. 3.4H**).

The population of ~50 nt sub-disome RPFs was unexpected. In most cases, sub-disome RPFs had 3′ ends 14 – 21 nt downstream of histidine codons, similar to monosome and true disome RPFs, but the 5′ ends were truncated 6 – 24 nt relative to true disome RPFs (**Fig. 3.4G**). RNase I digestion is known to produce both a ~58 nt true disome RPF and a 3′-truncated ~51 nt RPF with cleavage at the unoccupied A-site of the leading stalled ribosome (Meydan and Guydosh 2020). P1 nuclease did not appear to cleave within the unoccupied A-site of a stalled ribosome, however. Sub-disome RPFs with variable 5′ truncation could arise from cleavage of mRNA partially protected at an intermediate stage of trailing (upstream) ribosome removal by RQC (**Fig. 3.4G,** depicted in the lower panel). Such intermediates might accumulate when RQC pathways are saturated by collisions following Hts1 depletion. We also noted a less abundant class of sub-disome RPFs with 5′ ends corresponding to monosomes stalled on the histidine codon and a broad distribution of 3′ ends downstream of this stall site.

Sub-disome RPFs were also enriched around start codons, with few true disomes at these positions (**Fig. 3.4I**). Similar to the profile at histidine codon stalls, the majority of sub-disomes at start codons had 5′ ends ranging from 18 – 40 nt upstream of the start codon and 3′ ends 18 – 24 nt downstream of the start codon, similar to the 3′ ends of RPFs corresponding to initiating monosomes (**Fig. 3.4I**, **Extended Data 3.4E-F**). A minor population of sub-disome RPFs had 5′ ends 13 – 14 nt upstream of start codon, in common with initiating monosomes (**Extended Data 3.4E-F**), but a broader distribution of 3′ ends (**Fig. 3.4I,** see double-asterisks), paralleling the minor population of sub-disome RPFs at histidine codon stalls (**Fig. 3.4G,** see double-asterisks). We also captured true disome footprints consistent with a leading ribosome positioned at the stop codon (**Fig. 3.4J**). Considering these results overall, we propose that P1 nuclease digestion preserves larger RPFs around sites of translation initiation and elongation stalling, perhaps due to specialized factors that associate with the ribosome at these positions. These results highlight the utility of P1 nuclease digestion for mapping ribosome collisions and other higher-order ribosome complexes.

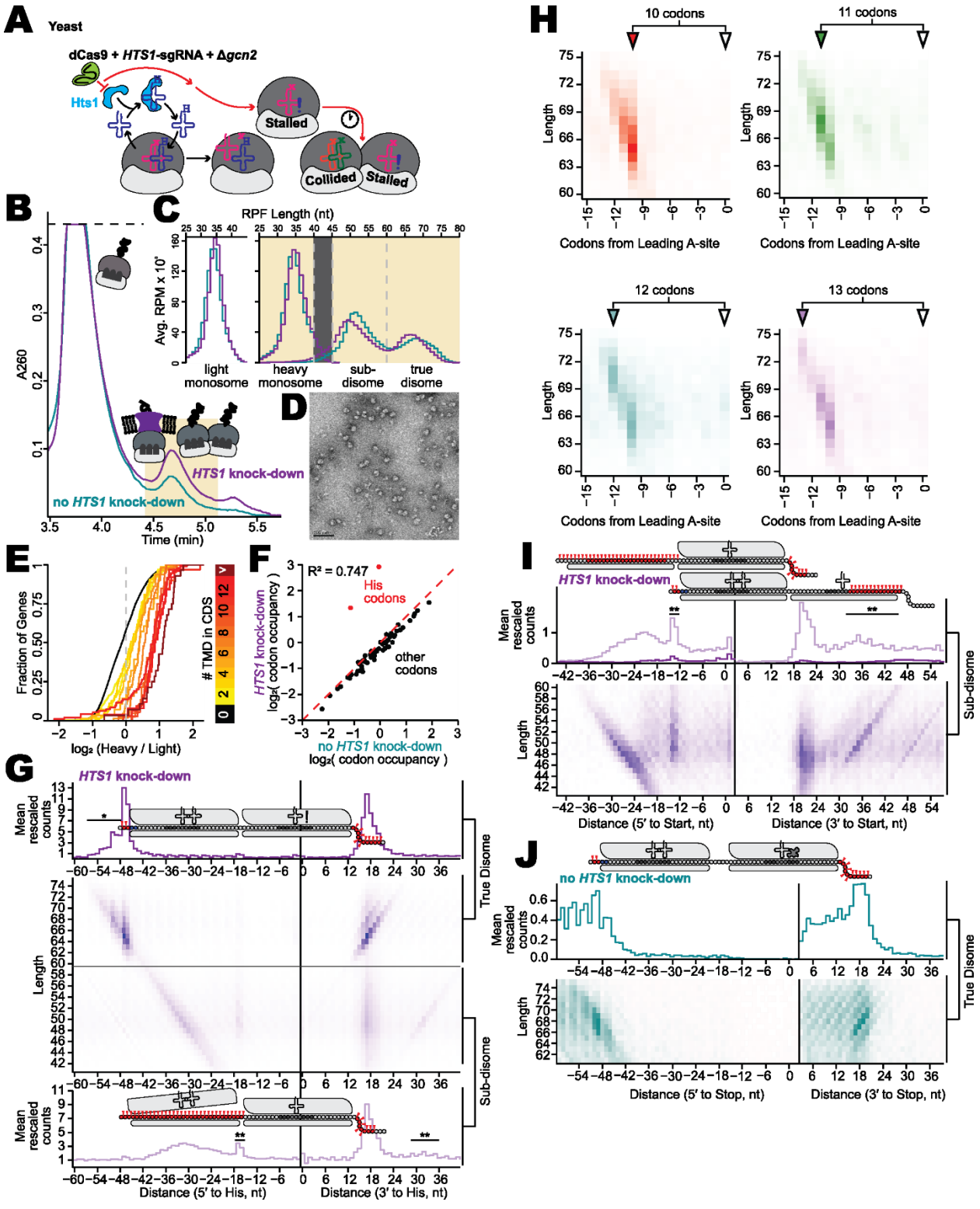

A
Yeast
dCas9 + HTS1-sgRNA + Δgcn2
Hts1
Stalled
Collided
Stalled
B
A260
Time (min)
HTS1 knock-down
no HTS1 knock-down
C
RPF Length (nt)
Avg. RPM x 10^3
light monosome
heavy monosome
sub-disome
true disome
D
E
Fraction of Genes
log2 (Heavy / Light)
TMD in CDS
F
R² = 0.747
His codons
other codons
HTS1 knock-down log2(codon occupancy)
no HTS1 knock-down log2(codon occupancy)
G
HTS1 knock-down
Mean rescaled counts
Length
True Disome
Sub-disome
Distance (5′ to His, nt)
Distance (3′ to His, nt)
H
10 codons
11 codons
12 codons
13 codons
Length
Codons from Leading A-site
I
HTS1 knock-down
Mean rescaled counts
Length
Sub-disome
Distance (5′ to Start, nt)
Distance (3′ to Start, nt)
J
no HTS1 knock-down
Mean rescaled counts
Length
True Disome
Distance (5′ to Stop, nt)
Distance (3′ to Stop, nt)

Figure 3.4: P1 nuclease disome and sub-disome footprints.

A, Schematic of translatome-wide histidine codon stalling by knockdown of *HTS1*, encoding histidyl-tRNA synthetase.

B, Sucrose density gradient profile of P1 nuclease digested yeast lysates with or without *HTS1* knockdown. Yellow shaded box indicates the disome fraction. One of two replicates shown.

C, RPF length distribution from light monosome, heavy monosome, and disome libraries, as in **Extended Data 3.1B**. The dashed line indicates the separation between sub-disome and true disome lengths.

D, A representative negative stain electron micrograph of a disome fraction from P1 nuclease digested yeast lysate, from two independent samples and 84 micrographs. Scale bar: 100 nm.

E, Comparison of light and heavy monosome RPF occupancy. Cumulative distribution of enrichment ratios were computed for genes stratified by the number of transmembrane domains (TMD) they encode.

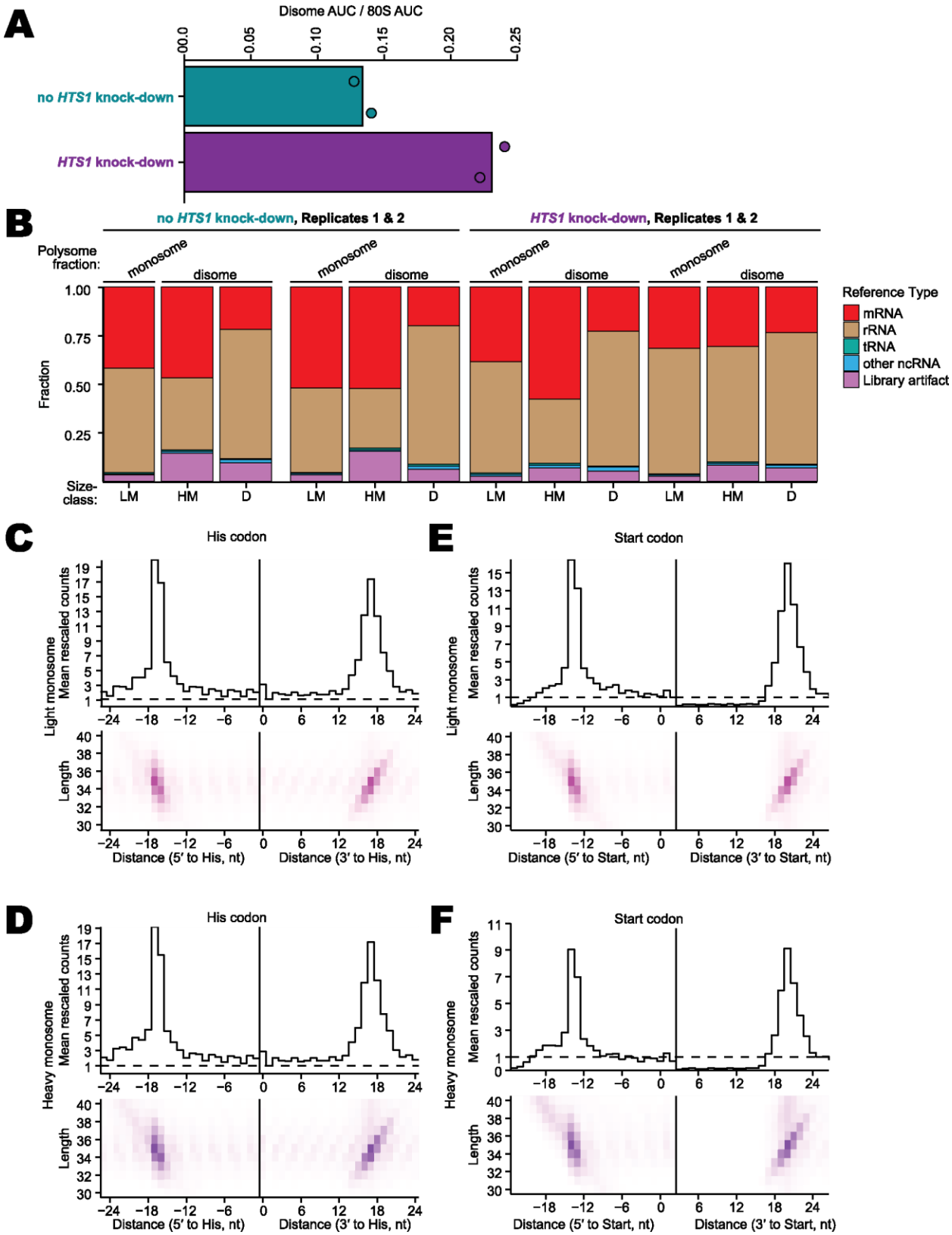
F, Comparison of codon-level ribosome occupancy estimates, as in **Fig. 3.1C**, from light-monosome libraries with or without *HTS1* knockdown. Pearson's R^2 is reported for all codons including histidine.

G, As in **Fig. 3.3E**, profile of true disome (60 – 75 nt) and sub-disome (40 – 59 nt) RPFs around isolated histidine codons after *HTS1* knockdown. Collisions with leading ribosomes at histidine codons are schematized; on the bottom, the trailing ribosome is undergoing RQC. The 60S (top) and 40S (bottom) subunits are shown with P- and A-site tRNAs, or an exclamation point for missing histidyl-tRNA, enclosing an mRNA (circles). P1 cleavage positions denoted by red triangles.

H, Profile of true disome RPFs in *HTS1* knockdown where pairs of histidine codons occur with 10 – 13 intervening non-histidine codons.

I, As in **Fig. 3.3E**, start site profile of sub-disome and true disome RPFs in *HTS1* knockdown, with an initiating ribosome and upstream pre-initiation complex schematized.

J, As in **Fig. 3E**, termination profile of true disome footprint read lengths without *HTS1* knockdown, with nearly-collided ribosomes at termination schematized.


A
Disome AUC / 80S AUC
0.0
0.05
0.10
0.15
0.20
0.25
no HTS1 knock-down
HTS1 knock-down
B
no HTS1 knock-down, Replicates 1 & 2
HTS1 knock-down, Replicates 1 & 2
Polysome fraction:
monosome
disome
Fraction
1.00
0.75
0.50
0.25
Size-class:
LM
HM
D
Reference Type
mRNA
rRNA
tRNA
other ncRNA
Library artifact
C
His codon
Light monosome
Mean rescaled counts
Length
Distance (5′ to His, nt)
Distance (3′ to His, nt)
E
Start codon
Distance (5′ to Start, nt)
Distance (3′ to Start, nt)
D
Heavy monosome
F

Extended Data 3.4: Abundance and properties of monosome, sub-disome, and true disome RPFs from yeast cells with or without histidine starvation.

A, Ratio of sucrose density gradient disome to monosome peak area from polysome profiles after P1 nuclease digestion of yeast lysates from cells with or without *HTS1* knockdown.

B, Fraction of sequencing reads mapped to each transcript class for sucrose density gradient light monosome, heavy monosome, and disome profiling from P1 nuclease digestion and OTTR library cDNA synthesis, as in **Extended Data 3.1A**.

C, Average profile of sucrose density gradient light monosome RPFs around isolated histidine codons after *HTS1* knockdown, as in **Fig. 4G.**

D, Average profile of sucrose density gradient heavy monosome RPFs around isolated histidine codons after *HTS1* knockdown, as in **Fig. 4G.**

E, Average profile of sucrose density gradient light monosome RPFs around start codons after *HTS1* knockdown, as in **Fig. 3.3E.**

F, Average profile of sucrose density gradient heavy monosome RPFs around start codons after *HTS1* knockdown, as in **Fig. 3.3E.**

Distinctive P1 nuclease footprints from uORF translation

In addition to ribosome stalling at histidine codons, Hts1 depletion triggers the integrated stress response through translational up-regulation of general control nondepressible 4 (*GCN4*) (Hinnebusch 2005). We recently reported that this induction is *GCN2* independent, and so we expected to see elevated *GCN4* translation (Muller et al. 2020). Indeed, we saw roughly 2-fold higher ribosome occupancy on the *GCN4* CDS after *HTS1* knockdown (**Fig. 3.5A**) (Chothani et al. 2019).

Translational control of *GCN4* expression depends on short upstream ORFs (uORFs) in its 5′ leader. Previous ribosome profiling using RNase I suggested that most ribosome occupancy on the *GCN4* 5′ leader occurs over the first, 3-codon uORF (uORF1) (Ingolia et al. 2009). We found the same pattern for monosome-sized P1 nuclease RPFs with or without *HTS1* knockdown, although some footprints were lengthened slightly on the 5′ end (**Fig. 3.5B-C, Extended Data 3.5A-D**). Upon *HTS1* knockdown, sub-disome and true disome sized footprints extending upstream of the monosome RPFs could be seen, reminiscent of extended 40S footprints observed at this uORF (Young et al., 2021) (**Fig. 3.5B, Extended Data 3.5A-D**). In contrast to uORF1, sub-disome and true disome footprints were abundant at the second, 2-codon uORF (uORF2) under both conditions, while monosomes were sparse (**Fig. 3.5B, Fig. 3.5D, Extended Data 3.5E-H**).

The *CPA1* transcript leader harbors a longer regulatory uORF that encodes the arginine attenuator peptide (AAP), which stalls ribosomes at the terminating codon when arginine is abundant (Meydan and Guydosh 2020, Wei et al. 2012, Gaba et al. 2021). We observed many true disome RPFs produced when the leading ribosome was paused over the uORF stop codon (**Extended Data 3.5I**). Surprisingly, these footprints were longer (**Extended Data 3.5J**) than collided disomes on histidine codons after *HTS1* knockdown (63 – 65 nt) (**Fig. 3.4G**). Fortuitously, the last codon of the *CPA1* uORF encodes histidine, and *HTS1* knockdown shifted the 3′ ends of uORF disome footprints one codon upstream, corresponding with the leading ribosome paused at the histidine codon, without altering the 5′ end positions (**Extended Data 3.5I-J, Extended Data 3.5K-L**).

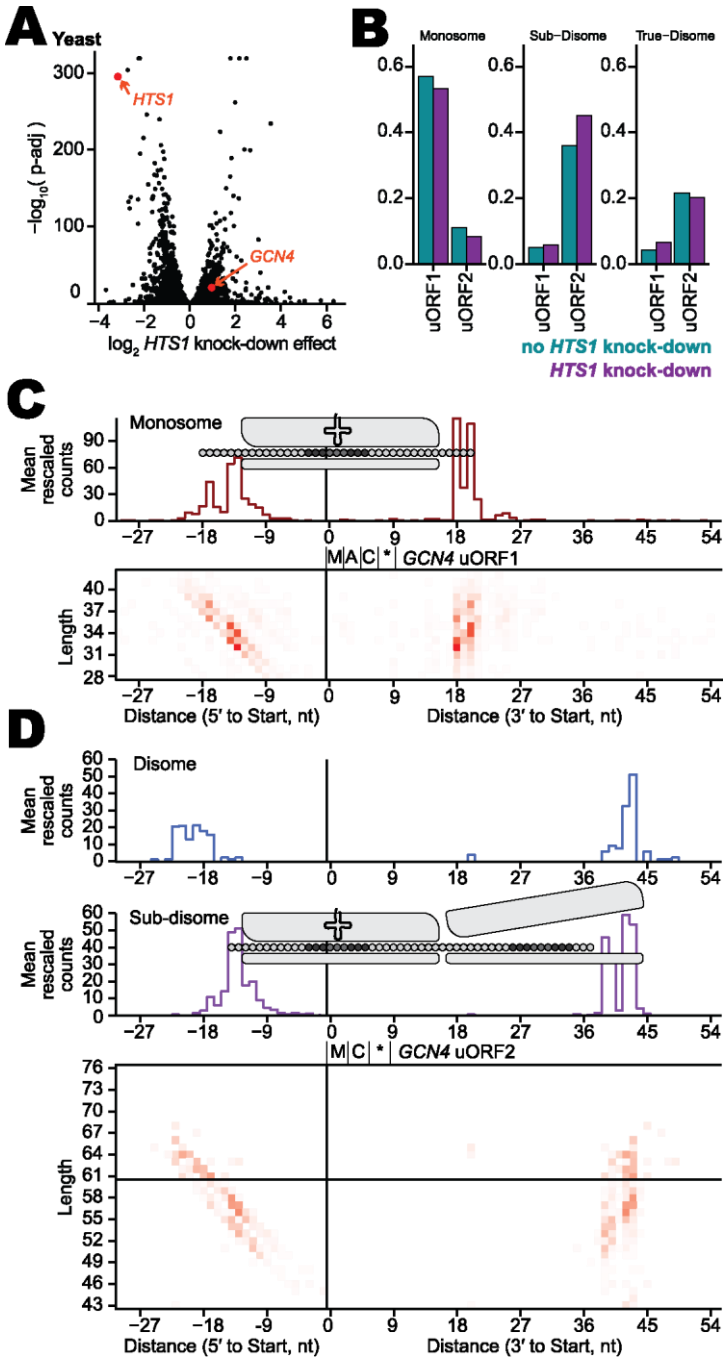

A
Yeast
HTS1
GCN4
−log10(p-adj)
log2 HTS1 knock-down effect
B
Monosome
Sub−Disome
True−Disome
uORF1
uORF2
no HTS1 knock-down
HTS1 knock-down
C
Monosome
Mean rescaled counts
M|A|C|*| GCN4 uORF1
Length
Distance (5′ to Start, nt)
Distance (3′ to Start, nt)
D
Disome
Sub-disome
M|C|*| GCN4 uORF2

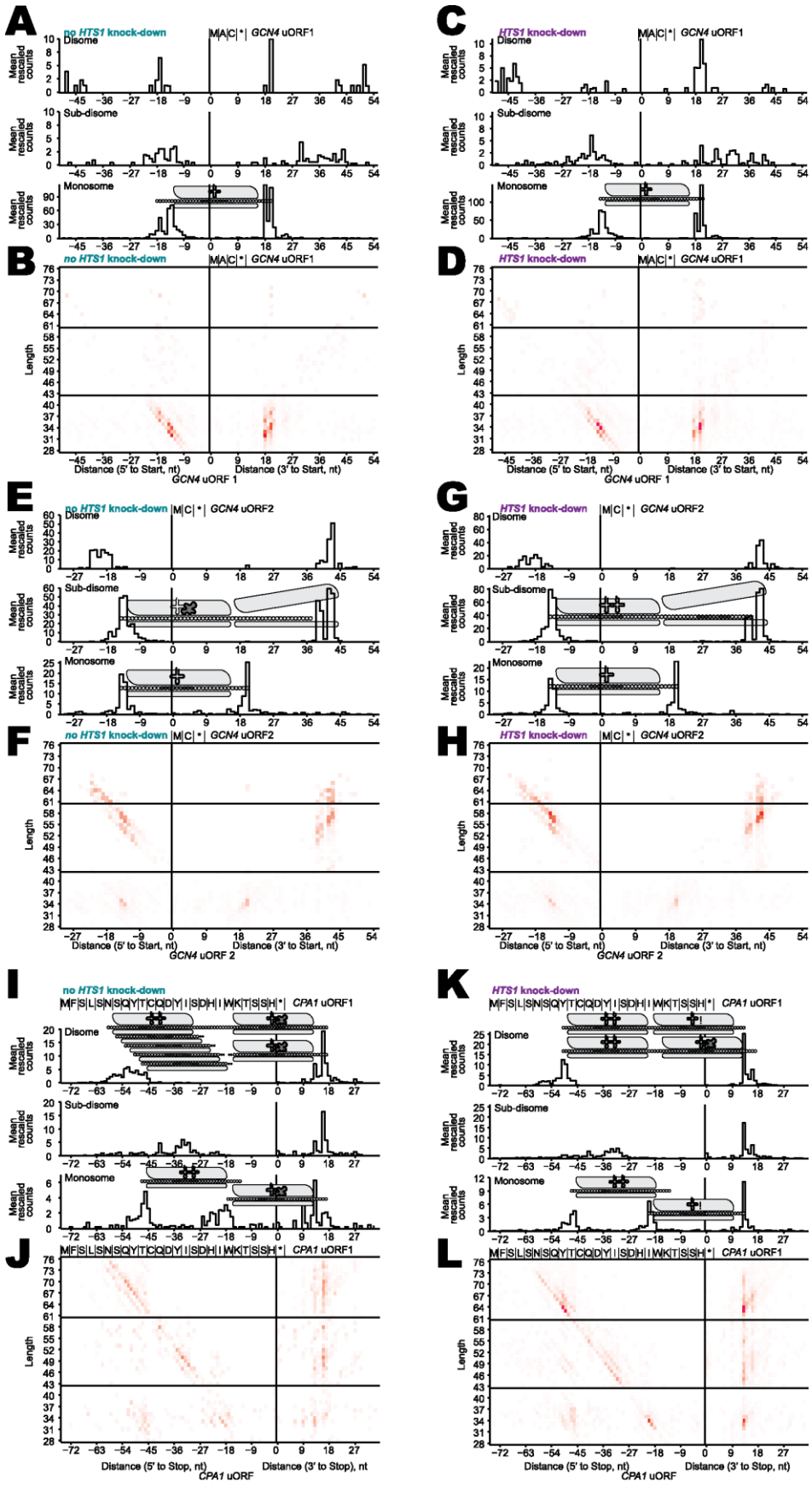
Figure 3.5: Monosome, sub-disome, and true disome occupancy profiles for *GCN4* uORF1 and uORF2 and *CPA1* uORF.

A, Differential expression analysis of sucrose density gradient monosome footprint occupancy across yeast genes, with versus without *HTS1* knockdown. *HTS1* and *GCN4* are shown in red. P-values from Wald test were adjusted for multiple testing using the Benjamini-Hochberg method.

B, Average fraction of aligned footprints to uORF1 and uORF2 of *GCN4* from sucrose density gradient purified light monosomes, sub-disome, and true disomes. Counted alignments represented as fractions for each library before replicate libraries were averaged. Alignments were assigned to their respective ORF if both the 5′ and 3′ ends flanked the start and stop codons of the uORF but did not overlap another uORF.

C, Sucrose density gradient monosome 5′ and 3′ profile of ribosomes at *GCN4* uORF1. Alignments were rescaled by the mean CDS codon occupancy of *GCN4* before replicates were summed together.

D, Sucrose density gradient true disome and sub-disome 5′ and 3′ profile of ribosomes at *GCN4* uORF2. Disome and sub-disome rescaled counts were normalized separately. Alignments for disome and sub-disome RPFs were rescaled by the respective mean codon count of *GCN4*'s CDS, excluding those RPFs that are aligned to the first 15 and last 10 codons, before replicates were combined by summing together.


A
no HTS1 knock-down
Disome
Sub-disome
Monosome
Mean rescaled counts
M|A|C|*| GCN4 uORF1
B
no HTS1 knock-down
Length
Distance (5' to Start, nt)
Distance (3' to Start, nt)
GCN4 uORF 1
C
HTS1 knock-down
D
HTS1 knock-down
E
no HTS1 knock-down
M|C|*| GCN4 uORF2
F
no HTS1 knock-down
GCN4 uORF 2
G
HTS1 knock-down
H
HTS1 knock-down
I
no HTS1 knock-down
M|F|S|L|S|N|S|Q|Y|T|C|Q|D|Y|I|S|D|H|I|W|K|T|S|S|H|*| CPA1 uORF1
Distance (5' to Stop, nt)
Distance (3' to Stop), nt
CPA1 uORF
J
M|F|S|L|S|N|S|Q|Y|T|C|Q|D|Y|I|S|D|H|I|W|K|T|S|S|H|*| CPA1 uORF1
K
HTS1 knock-down
L
HTS1 knock-down

Extended Data 3.5: Complete *GCN4* uORF1 and uORF2 and *CPA1* uORF profiles for monosome, sub-disome, and true disome footprints with and without *HTS1* knockdown.

A, Extended representation from **Fig. 3.5C**. Rescaled counts of 5′ and 3′ ends of aligned true disome (top), sub-disome (middle), and monosome (bottom) footprints for *GCN4* uORF1 without *HTS1* knockdown. Results from replicates were summed together from sucrose density gradient purified material.

B, Footprint length 5′ and 3′ profile of (**A**).

C, Rescaled counts of 5′ and 3′ ends of aligned true disome (top), sub-disome (middle), and monosome (bottom) footprints for *GCN4* uORF1 after *HTS1* knockdown. Results from replicates were summed together.

D, Footprint length 5′ and 3′ profile of (**C**).

E, Extended representation from **Fig. 3.5D**. Rescaled counts of 5′ and 3′ ends of aligned true disome (top), sub-disome (middle), and monosome (bottom) footprints for *GCN4* uORF2 without *HTS1* knockdown. Results from replicates were summed together.

F, Footprint length 5′ and 3′ profile of (**E**).

G, Rescaled counts of 5′ and 3′ ends of aligned true disome (top), sub-disome (middle), and monosome (bottom) footprints for *GCN4* uORF2 after *HTS1* knockdown. Results from replicates were summed together.

H, Footprint length 5′ and 3′ profile of (**G**).

I, Rescaled counts of 5′ and 3′ ends of aligned true disome (top), sub-disome (middle), and monosome (bottom) footprints for *CPA1* uORF without *HTS1* knockdown. Results from replicates were summed together. Both a terminating ribosome (leading) and elongating ribosome (trailing) at various codon positions are schematized. Results from replicates were summed together.

J, Footprint length 5′ and 3′ profile of (**I**).

K, Rescaled counts of 5′ and 3′ ends of aligned true disome (top), sub-disome (middle), and monosome (bottom) footprints for *CPA1* uORF with *HTS1* knockdown. Both an elongating ribosome stalled at the penultimate histidine codon (leading) and an elongating ribosome (trailing) are schematized. Results from replicates were summed together.

I, Footprint length 5′ and 3′ profile of (**K**).

Ribosome profiling without gel purification of RPFs

One remaining impediment to high-throughput ribosome profiling is the use of gel-based RPF size selection, needed to eliminate rRNA fragments produced by RNase I. Given the limited rRNA fragmentation by P1 nuclease (**Fig. 3.2E**), we wanted to test whether gel-based RPF selection could be replaced by mirRICH small RNA enrichment, which relies on differential resuspension of small (< 200 –300 nt) versus large RNAs (Choi et al. 2018). We carried out an independent P1 nuclease digestion of the Δ*gcn2* yeast lysates described above (**Fig. 3.4 – 6**), recovered ribosomes using a sucrose cushion, extracted RNA, and split the material in half, using either total RNA or mirRICH small RNA as input for library synthesis and performing size selection only after OTTR cDNA synthesis and after library amplification (**Extended Data 3.6A-C**). Gene-level quantifications of ribosome occupancy using mirRICH or total RNA monosome libraries closely matched results from gel-based size selection (**Fig. 3.6A, Extended Data 3.6D**), however the total RNA (but not mirRICH) results were distorted by PCR duplicates (**Extended Data 3.6E**).

To assess whether monosome and disome libraries could be generated from the same OTTR library of mirRICH RPFs, we selected cDNA to individually isolate monosome- and disome-sized inserts. Indeed, cDNA size selection sufficed to separate disome- from monosome-sized reads (**Extended Data 3.6F**), and we captured sub-disome footprints upstream of start sites (**Fig. 3.6B**) and true disome footprints from stalled ribosomes (**Fig. 3.6C**). Without fractionation, we did see extended footprints at mRNA positions expected to have stalled monosomes (**Fig. 3.6C,** RPF ends indicated by asterisks), perhaps because P1 nuclease failed to fully degrade the unprotected mRNA 3′ of a ribosome.

We next applied mirRICH enrichment and cDNA size selection to perform monosome and disome profiling from 293T human cell lysate. As in yeast, we performed P1 nuclease digestion, ribosome isolation and OTTR cDNA synthesis, followed by gel purification of monosome and disome cDNAs (**Extended Data 3.6G**). We again observed highly correlated ribosome occupancy measurements between gel-purified and mirRICH monosome RPFs (**Fig. 3.6D**). We then examined libraries representing 50 – 80 nt RNAs, which should contain human disome and sub-disome footprints. Sub-disome RPFs accumulated around the sites of translation initiation, as noted in yeast, with few true-disome sized RPFs appearing before codons ~15 – 18 (**Extended Data 3.6H**). At termination, we noted a dramatic enrichment of true-disome sized RPFs, including extended footprints from nearly-collided ribosomes, consistent with a leading ribosome on the stop codon (**Fig. 3.6E**).

We investigated the distribution of human disomes to identify potential site-specific stalling. Looking across well-translated CDSs, we identified disomes (>58 nt) with trailing A-site positions at *RPS7* (codon Glu14), *TUBA1A* (Tyr312), *CCT5* (Arg24), *HSPE1* (STOP103), and *ACTB* (Arg147). The major disome in *TUBA1A,* which encodes α-tubulin, was well separated from initiation or termination codons (**Fig. 3.6F**). Two discrete sites of disome accumulation were evident, with trailing ribosomes at the codons for Tyr312 and Tyr357 (**Fig. 3.6F-G**), along with monosome-sized footprints

corresponding to the leading and trailing ribosomes (**Fig. 3.6H**). Notably, at a homologous position of yeast *TUB1* (amino acids 355 – 368 in *TUB1* corresponding to amino acids 354 – 367 in *TUBA1A*), slow translation was associated with the recruitment of the TRiC/CCT chaperonin complex (Stein et al. 2019). Disome accumulation at these positions could be explained by a similar phenomenon (Zhao et al. 2021).

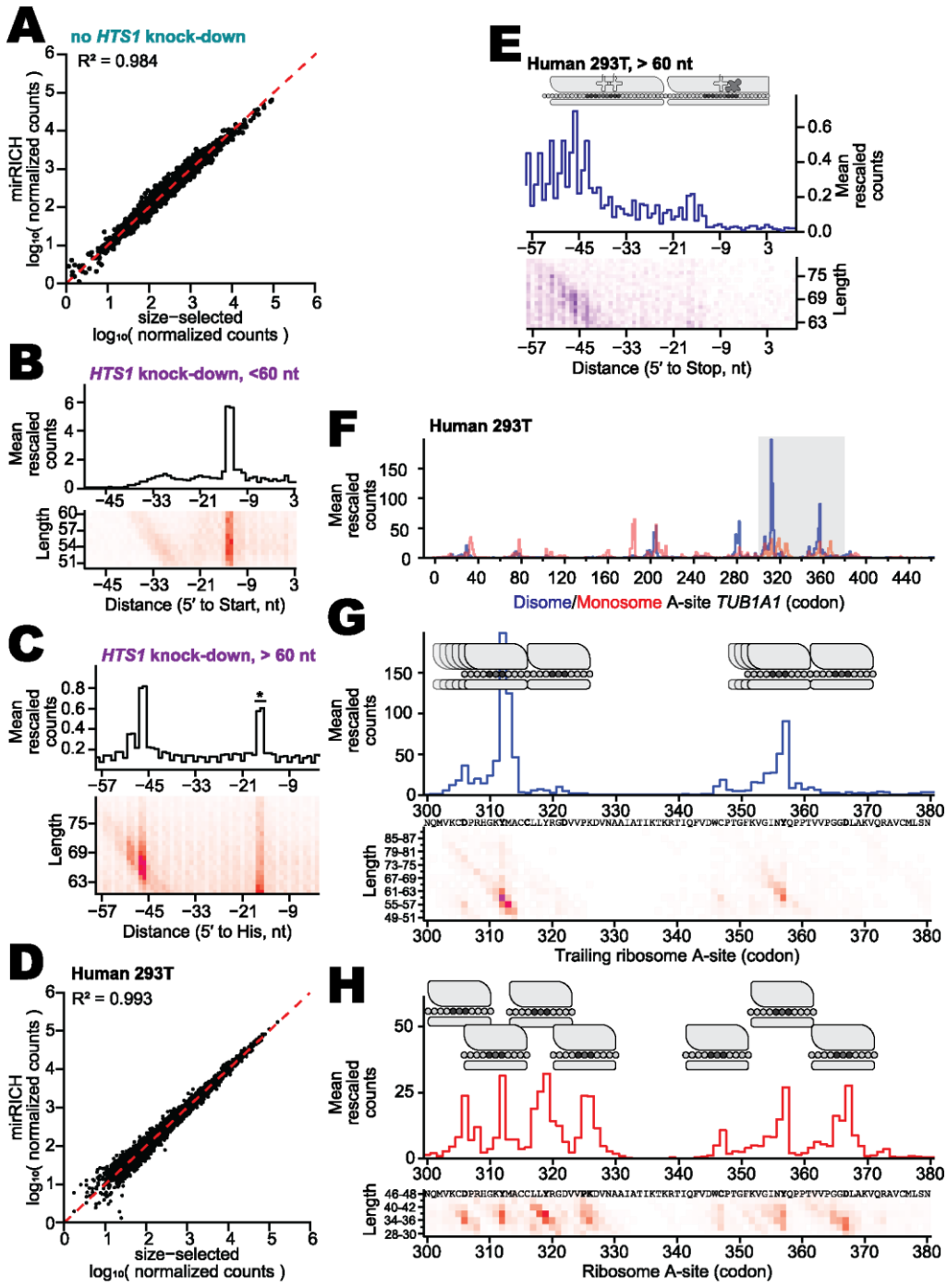


Figure 3.6: mirRICH and cDNA size selection as alternatives to gel-based size selection of P1 nuclease RPFs.

A, Gene-level estimates for CDS occupancy comparing sucrose cushioned and mirRICH purified RPFs to sucrose density gradient and gel-based size selected light-monosome RPF from yeast lysate without *HTS1* knockdown. mirRICH relied on cDNA size selection whereas light-monosome relied on polysome fractionation followed by both RNA size selection and cDNA size selection. Read counts for each gene were normalized by DESeq2.

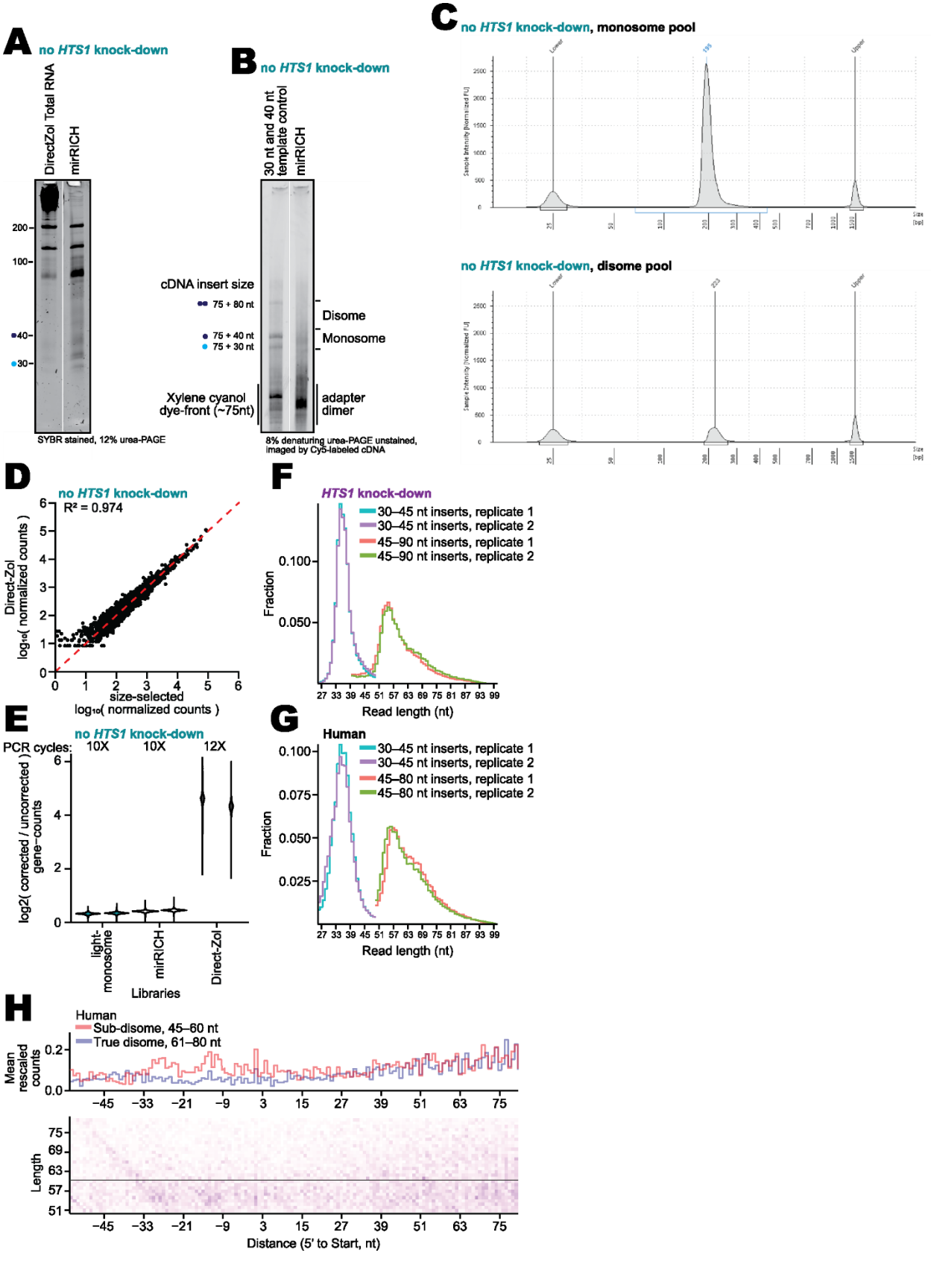
B, C, Footprint 5′ ends and length profile at initiation codons for (**B**) P1 sub-disome RPFs and (**C**) P1 true disome RPFs, captured by sucrose cushioned and mirRICH from *HTS1* knockdown yeast lysate. The summed 5′ end profile is depicted at top and the contributions from each footprint length at each 5′ end are shown in the matrix at bottom.

D, Gene-level estimates for CDS occupancy comparing mirRICH purified RPFs to gel-based size selection for human 293T P1 RPFs. Material was purified by a sucrose cushion. Read counts for each gene were normalized by DESeq2.

E, Footprint 5′ ends and length profile at termination codons for P1 true disome RPFs captured by mirRICH from human 293T cell sucrose cushion purified material. The summed 5′ end profile is depicted at top and the contributions from each footprint length at each 5′ end are shown in the matrix at bottom.

F, Occupancy profile for sucrose cushion purified monosomes (red) and trailing ribosome of the disome (blue) across the human *TUB1A1* ORF. The region in gray is expanded in **Fig. 3.6G-H**.

G-H, Profile of A-site positions of sucrose cushion purified (**G**) trailing ribosomes from true disome footprints and (**H**) monosome footprints, with length and position grouped by codon, from the gray region in (**F**). Monosomes (**H**) and disomes (**G**) are schematized above to emphasize differences in leading and trailing ribosome dwell time.


A
no HTS1 knock-down
DirectZol Total RNA
mirRICH
200
100
40
30
SYBR stained, 12% urea-PAGE
B
no HTS1 knock-down
30 nt and 40 nt template control
mirRICH
cDNA insert size
75 + 80 nt
75 + 40 nt
75 + 30 nt
Xylene cyanol dye-front (~75nt)
Disome
Monosome
adapter dimer
8% denaturing urea-PAGE unstained, imaged by Cy5-labeled cDNA
C
no HTS1 knock-down, monosome pool
no HTS1 knock-down, disome pool
D
no HTS1 knock-down
R² = 0.974
Direct-Zol log10(normalized counts)
size-selected log10(normalized counts)
E
no HTS1 knock-down
PCR cycles: 10X 10X 12X
log2(corrected / uncorrected) gene-counts
light-monosome
mirRICH
Direct-Zol
Libraries
F
HTS1 knock-down
30–45 nt inserts, replicate 1
30–45 nt inserts, replicate 2
45–90 nt inserts, replicate 1
45–90 nt inserts, replicate 2
Fraction
Read length (nt)
G
Human
30–45 nt inserts, replicate 1
30–45 nt inserts, replicate 2
45–80 nt inserts, replicate 1
45–80 nt inserts, replicate 2
Fraction
Read length (nt)
H
Human
Sub-disome, 45–60 nt
True disome, 61–80 nt
Mean rescaled counts
Length
Distance (5' to Start, nt)

Extended Data 3.6: Additional comparisons of libraries made from mirRICH, total RNA, or gel-base size-selected RPFs.

A, A SYBR Gold stained 0.6X TBE 12% denaturing urea-PAGE gel of resolved P1 nuclease digested yeast RNA purified from either Direct-Zol Total RNA extraction or mirRICH small RNA enrichment following purification by sucrose cushion. Roughly 5% of the Direct-Zol extract and 25% of the mirRICH extract was analyzed in this gel. The migration of the 30 nt (light blue) and 40 nt (dark blue) size selection RNA oligos are demarcated (left). Both lanes are from the same gel, and both purification methods were replicated twice.

B, As in **Extended Data 3.3A**, cDNA size selection from a OTTR library generated from 40 ng of mirRICH small RNA from (**A**). Parallel positive control OTTR libraries were synthesized from either 30 or 40 nt size selection RNA oligos to aid in cDNA size selection. The single light-blue and dark-blue dots (left) demarcate the cDNA with inserts derived from the 30 nt and 40 nt RNA oligos, respectively. cDNA concatemers from deliberately using more templates in OTTR, e.g. ~80 nt insert, are demarcated by two dark blue dots. Monosome cDNA size selection area was demarcated by the bottom two black lines (right), and the disome cDNA size selection by the top two black lines. Both lanes are from the same gel. Four yeast and two human libraries from mirRICH purified RNA were constructed.

C, Library length distribution of bifurcated monosome (top) and disome (bottom) libraries by Agilent 2200 TapeStation. Libraries were bifurcated at the cDNA size selection step as shown in **Extended Data 3.6B.**

D, Gene-level estimates for CDS occupancy, as in **Fig. 3.1A**, but here for mirRICH or Direct-Zol P1 nuclease RPF from the yeast lysate without *HTS1* knockdown, purified by sucrose cushion. Both libraries relied on cDNA size selection rather than RNA size selection. The counts were deduplicated before analysis.

E, Distribution of the $\log_2$ ratio of PCR deduplicated counts (corrected) versus uncorrected counts for each CDS from the libraries generated from the no *HTS1* knockdown lysate, purified by sucrose cushion. The number of necessary PCR cycles used for library multiplexing is defined above. Two technical replicates per library condition were analyzed.

F, Read length distribution represented as a fraction of mRNA mapping reads for the sucrose cushion purified mirRICH monosome and disome reads bifurcated by cDNA size-selected libraries from the *HTS1* knockdown lysate. Two technical replicates per library condition were analyzed. Monosome replicates are in light blue and purple; disome replicates are in red and green.

G, Read length distribution represented as a fraction of mRNA mapping reads for the sucrose cushion purified mirRICH monosome and disome reads bifurcated by cDNA size-selected libraries from the 293T lysate. Two technical replicates per library

condition were analyzed. Monosome replicates are in light blue and purple; disome replicates are in red and green.

H, 5′ aligned ends and footprint length profile at initiating codons for P1 sub-disome (red line) and true disome (blue line) RPFs captured by mirRICH from human 293T cell lysates. Material was purified by a sucrose cushion. The summed 5′ end profile is depicted at top and the contributions from each footprint length at each 5′ end are shown at bottom. Sub-disome and true disome sized reads were analyzed separately.

Discussion

P1 nuclease digestion, mirRICH enrichment of RPFs, and OTTR cDNA library synthesis together provide a streamlined approach to ribosome profiling that offers many experimental advantages as well as increased biological insight from distinct classes of RPFs from apposed ribosomes. OTTR library generation provides adapter incorporation and reverse transcription simultaneously in a single-tube, 4-hour workflow with low input RPF requirements and low bias in capturing RNAs with different terminal nucleotides. We also demonstrate that P1 nuclease digestion improves monosome and disome yields relative to RNase I, reduces unwanted rRNA fragment contamination, and eliminates the need for end repair. Footprints produced by P1 nuclease digestion are larger than those from RNase I, enabling improved precision of sequence mapping to the genome. Analyses of P1 footprinting data must account for the range of P1 footprint sizes and should infer the positions of ribosomes based on the 5′ end of the footprint, which is less variable. In addition, replacing gel-based size selection with small RNA enrichment further simplifies RPF library generation. The simplicity, efficiency, and sensitivity achieved by combining P1 nuclease, mirRICH, OTTR, and Cy5-labeled cDNA eliminates days of labor and dozens of steps, providing the simplest ribosome profiling approach to date and suggesting a route towards automation. Moreover, these improvements may translate to a broader adoption of ribosome profiling over either mRNA sequencing or proteomic mass spectroscopy to access cell state.

Because P1 nuclease preserves higher-order ribosome complexes, it offers additional insights about mechanisms governing normal and aberrant translation. We did not see evidence of P1 nuclease cleavage of mRNAs within empty A-sites, in contrast to results from RNase I digestion, and so analyses that rely on these 21 nt footprints are probably not possible with P1 (Wu et al. 2019, Meydan and Guydosh 2020, Zhao et al. 2021). We did find that disomes resistant to P1 nuclease digestion include both collided ribosomes protecting a ~65 nt footprint and larger disome footprints from nearby but not collided pairs of ribosomes.

The sub-disome RPFs preserved by P1 nuclease have not been previously investigated. They accumulate at start codons, stop codons, translation stalls, and uORFs. We suggest that they may derive from juxtaposition of an mRNA-engaged 40S subunit and 80S ribosome, with heterogeneous P1 nuclease cleavage spread across the 40S-bound mRNA region. At stall sites this 40S subunit may arise as a RQC

intermediate that is pushed forward by a trailing ribosome; these may accumulate in our study when RQC is saturated by translatome-wide stalling. At translation start sites, the 40S may be a pre-initiation complex queued upstream of the initiating ribosome (Kainuma and Hershey 2001, Garshott et al. 2021). At termination, an unrecycled small subunit may be displaced into the 3′ untranslated region. Alternatively, sub-disome RPFs may instead reflect single ribosomes accompanied by accessory factors that extend the footprint. Selective profiling of P1 footprints will likely provide further insights. The sub-disome and true disome RPFs generated by P1 nuclease hold promise for deeper insights into translation kinetics and ribosome-associated RNA and protein quality control.

Materials and methods

Polysome buffers for cell lysis and polysome profiling.

For both yeast and human cell lysis, buffers and protocols described in McGlincy and Ingolia 2017, were followed. Polysome buffer was 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 5 mM MgCl2, and 1 mM DTT; for pH 6.5 polysome buffer, 20 mM Bis-Tris pH 6.5 was used instead of Tris-HCl pH 7.5. Lysis buffer included 1% (v/v) Triton X-100 and 25 U/mL Turbo DNase I. Buffers were made fresh and stored on ice; RNase-free commercial buffers were preferentially used, otherwise buffers were made in DEPC-treated water, filtered with a 0.22 µm pore, and when possible, autoclaved.

Yeast culture and harvesting for polysome and ribosome profiling

Fig. 3.1 – 3 and Extended Data 3.1 – 3 used BY4741 yeast. Fig. 3.4 – 6 and Extended Data 3.4 – 6 used NIY451 (BY4741 Δ*gcn2::hphMX dCas9-Mxi1::TetR::kanMXsyn P(RPR1)Tet-HTS1sgRNA::KILEU2*) (Muller et al. 2020). Unless otherwise stated, yeast cells were cultured in sterilized yeast extract and peptone media supplemented with 2% (v/v) dextrose (YEPD) at 30 °C. Yeast cells were back-diluted to an optical density (OD_{600nm}) of 0.05 or lower and grown to OD_{600} of 0.5 – 0.6, rapidly filtered, submerged in liquid nitrogen, and combined with polysome lysis buffer before the frozen mass was subjected to cryogrinding (McGlincy and Ingolia 2017). Pulverized yeast cells were thawed on ice and clarified by centrifugation at 20,000 × g for 30 minutes at 4 °C. The supernatant was aliquoted into pre-chilled tubes and snap-frozen in liquid nitrogen before long-term storage at -80 °C. Cycloheximide was not used in this study.

Human adherent cell culture and harvesting for polysome and ribosome profiling

Human Calu-3 cell lysate was used in Figure 3.2 and Extended Data 3.2 while 293T cell lysate was used in Figure 3.3, Figure 3.6, Extended Data 3.3, and Extended Data 3.6. Cell cultures were grown in 15 cm dishes and harvested at least 24 hours after passage

with no more than 70-80% confluency. Media was aspirated and cells were washed with 5 mL of ice-cold 1X phosphate buffered saline (PBS) and the culture dish was rested on a bed of lightly salted ice. PBS was thoroughly aspirated and 600 µL of polysome lysis buffer was added dropwise. Cells were detached by scrapping and the lysate was transferred to a pre-chilled tube and incubated 10 minutes on ice, gently pipetted to reduce viscosity, and clarified by centrifugation at 20,000 × g for 30 minutes at 4 °C. The supernatant was aliquoted into pre-chilled tubes and snap-frozen in liquid nitrogen before long-term storage at -80 °C. Cycloheximide was not used in this study.

RNA quantification of cell lysate

Through direct comparison with a standard curve of 0.0 – 1.0 ng/µL rRNA (ThermoFisher), the rRNA concentration of thawed lysates diluted in 1X TE buffer was quantified using the protocol for Quant-iT RiboGreen RNA kit (ThermoFisher). The assay volume was transferred to Minicell Borosilicate thin glass cuvette (Promega) and arbitrary fluorescence units were measured for each sample and the standard curve by the Glomax Multi Jr. Detection System (Promega) using the Fluorometer Blue module. A linear regression from the standard curve and dilution factor were used to estimate the RNA concentration in a given lysate.

RNase I digestion of yeast and human lysate.

E. coli RNase I (10 U/µL, Epicentre) digestion occurred at room temperature for 45 minutes while lysate gently rotated or mixed and then placed on ice. We digested 30 µg of yeast lysate and 15 µg of Calu-3 or 293T cell lysate with 0.5 U/µg of RNase I in a 200 – 300 µL volume. For ribosome profiling, the RNase I RPFs were size-selected from 26 – 34 nt after resolving on denaturing 15% urea-PAGE and dephosphorylated by T4 PNK, as described (McGlincy and Ingolia 2017).

P1 nuclease digestion of yeast and human lysate.

We adjusted lysate to pH 6.5 by adding 7 µL of 300 mM Bis-Tris pH 6.0 per 100 µL of lysate and digested yeast lysate with 10 – 30 U/µg of P1 nuclease at 30 °C, or human cell lysate with 15 – 45 U/µg of P1 nuclease at 37 °C. Lysate was digested for 60 minutes and placed on ice. For ribosome profiling, the P1 nuclease digested RPFs were size-selected from 30 – 40 nt after resolving on denaturing 15% urea-PAGE, although we now recommend up to 45 nt for human ribosome profiling studies.

Polysome profiling

Sucrose density gradients were made in 14 x 89 mm Seton Open-Top Polyclear centrifuge tubes (Seton Scientific) from 6 mL of polysome buffered 10% (w/v) D-sucrose and 6 mL of polysome buffered 50% D-sucrose solution using the Gradient Station (BioComp Instruments). Using a wide-bore pipette tip, 300 µL of untreated or nuclease treated lysate was overlaid on top of the gradient. Lysate was resolved through the gradient by ultracentrifugation on a SW-41 rotor for 3 hours at 159,667 × g (36,000 RPM) at 4 °C and fractionated using the Gradient Station, measuring A_{260nm} using the BioRad Econo UV Monitor (BioRad) to measure the polysome profile and Logger Lite (v1.9.4) to record the polysome profile. The monosome and disome fractions were collected based on A_{260nm} profiles into pre-chilled tubes and snap-frozen in liquid nitrogen before storage at -80C. RNase-free reagents and filter-tip pipettes were used for each step.

Sucrose-cushioned pelleting of ribosomes.

In a 13 x 56 mm polycarbonate thick wall tube (Beckman Coulter), 200 – 300 µL of nuclease digested lysate was gently overlayed onto 900 µL of pre-chilled 1 M sucrose in polysome buffer using a wide-bore tip. Sample tubes were ultra-centrifuged for an hour at 408,800 × g (100,000 RPM) in a pre-chilled 4 °C TLA 110 rotor, the sucrose cushion was removed, and the ribosome pellet was disrupted with a pipet tip while being resuspended with 30 µL of 1X polysome buffer or nuclease-free water. After 10 minutes of incubation on ice, 300 µL of TRIzol reagent (ThermoFisher) was added, and it was transferred to a pre-chilled tube, vigorously vortexed, and incubated for 10 minutes at room temperature, followed by Direct-Zol RNA purification, mirRICH, or long-term storage at -80 °C.

Direct-Zol RNA extraction and urea-PAGE RNA size selection of RNase I or P1 RPFs

Direct-zol RNA Purification (Zymo Research) was performed as described by the manufacturer, including on-column DNase I digestion. RNA was eluted in 100 µL of nuclease free water and precipitated at -80 °C for ≥ 3 hours with 1 µL of 15 mg/mL GlycoBlue coprecipitant (ThermoFisher), 10 µL of 3 M sodium acetate pH 5.5, and 300 µL of 100% ethanol. The precipitate collected after centrifuging for 30 minutes at 20,000 × g at 4 °C was subsequently washed with 1 mL of -20 °C pre-chilled 70% ethanol, vortexed, incubated at -20 °C for 15 minutes, and centrifuged for 15 minutes at 20,000 × g at 4 °C, followed by a second identical wash. After the second wash, the RNA pellet was resuspended in 5 µL of 0.025% (w/v) bromophenol blue, 0.025% (w/v) xylene cyanol, 5 mM pH 8.0 EDTA, and 95% formamide (2X RNA loading dye) and either resolved by denaturing urea-PAGE electrophoresis or stored at -80 °C.

Glassware and plasticware used for RNA gel electrophoresis were soaked in 10% bleach for 5 minutes before washed by COUNT-OFF™ Surface Cleaner

(PerkinElmer) and liberally rinsing in double-distilled water. An RNase-free solution of 0.6X TBE, 7M UltraPure™ urea (ThermoFisher), and 15% 19:1 acrylamide:bis-acrylamide (Ambion) was polymerized by adding 1% ammonium persulfate and 0.1% UltraPure™ TEMED (ThermoFisher) and cast between borosilicate glass planes, allowing at least 30 minutes for polymerization. Once cast, gels were placed in a cleaned electrophoresis chamber in RNase-free 0.6X TBE running buffer. The gel was pre-run at 150 V for 15 minutes. RNA samples were resuspended in 5 µL of 2X RNA loading dye and were denatured at 80 °C for 90 seconds and snap cooled on ice until loaded. Wells were washed before 5 µL of denatured RNA was loaded. Wells were skipped between samples as necessary. When specified, a pair of either previously described RNase I RPF markers (RNA, 26 and 34 nt) (Ingolia et al. 2009) or a pair of P1 nuclease RPF markers (**Table 3.1**) were included. Electrophoresis was performed at 500 V. Gels were untouched after disassembly and manipulated exclusively by Saran-wrap or single-use blades to limit RNase and cross-nucleic acid contamination. Gels were stained for 1 minute with SYBR Gold (ThermoFisher). Gels were either analyzed on a Typhoon Trio (Cytiva) or illuminated on a Dark Reader Blue Light Transilluminator (Clare Chemical Research). RNase I RPFs were excised with the aid of the 26 nt to 34 nt maker while P1 RPFs were excised aided by the 30 nt to 40 nt marker. P1 disome RPFs were excised from 45 – 80 nt.

RNA elution from urea-PAGE

Gel slices were ground using a pipette tip, 400 µl of elution buffer with 300 mM NaCl, 10 mM Tris-HCl pH 7.0, 1 mM EDTA, and 0.25% (v/v) SDS was added, and incubated overnight. Gel debris was excluded by centrifuging 10,000 × g for 5 minutes at room temperature and removing ~350 µL of eluate. The eluate was combined with 1 µL of 15 mg/mL GlycoBlue coprecipitant and three volumes of 100% ethanol, and incubated for ≥ 3 hours at -80 °C and then centrifuged for 30 minutes at 20,000 × g at 4 °C. The pellet was washed with 1 mL of -20 °C pre-chilled 70% ethanol, and incubated at -20 °C for 15 minutes, and centrifuged again for 15 minutes at 20,000 × g at 4 °C for 2 - 3 total washes, and the precipitate air dried briefly before resuspended in 10 – 15 µL of RNase-free water.

RNase I RPF T4 polynucleotide kinase (PNK) dephosphorylation.

We purified RPFs using Oligo Clean and Concentrate (Zymo Research) according to the manufacturer's instructions. For every 30 µg of treated lysate, we treated the RNase I RPFs with 5 U of T4 PNK (NEB, 10 U/µL) after RNA size selection (McGlincy and Ingolia 2017) and prepared 10X T4 PNK Buffer (700 mM Tris-HCl pH 7.5, 100 mM $MgCl_2$, 50 mM DTT) with fresh DTT on the day of use.

Monosome and disome profiling by P1 nuclease profiling

Lysate containing 150 µg of total RNA was diluted to 300 µL with pH 7.5 polysome buffer and supplemented with 21 uL of 300 mM Bis-Tris (pH 6.0). Digestion was performed by adding 10 U/µg of P1 nuclease and incubating at 30 °C for 1 hour. Monosome and disome fractions were collected from polysome profiles as described above. Fractions were thawed on ice, diluted to 4 mL, and concentrated using single-use, pre-washed, Amicon Ultra-4 concentrators (Amicon) by centrifugation for 5 min at 4,000 × g at 4 °C, followed by two additional passes of 4 mL of ice-cold polysome buffer. A final volume of 100 – 200 µL was recovered and either snap-frozen in liquid nitrogen and stored at -80 °C or immediately combined with 300 – 600 µL of TRIzol for Direct-Zol RNA extraction as described above. Extracted RNA was resolved by 15% denaturing urea-PAGE as described for pelleted ribosomes. Monosome footprints were collected RNA from 30 – 40 nt while sub-disome and true disome footprint (from the disome fractions) were collected both 45 – 80 nt. Monsomes footprints were also collected from the disome fraction from 30 – 40 nt.

Δ*gcn2 HTS*1-sgRNA inducible yeast culture, harvesting, and monosome and disome profiling.

Cultures of NIY451 (Muller et al. 2020) were back-diluted from an overnight culture to an OD_{600} of 0.025 in either 750 mL YEPD with 250 ng/ml of anhydrotetracycline (for *HTS1* knockdown) or 250 mL of YEPD. The uninduced culture was harvested at an OD of 0.60, at which time we also harvested the induced culture at an OD_{600} of 0.21.

OTTR ribosome profiling library generation.

OTTR library generation requires two recombinantly expressed, modified non-LTR retroelement R2 RT proteins, Relay Pol and Tail Pol (Upton et al. 2021, Pimentel et al. 2022). Relay Pol is a recombinantly expressed truncated non-LTR R2 RT with an MBP and 6x-His tag. Tail Pol is the same protein but includes the mutation F753A, which increases turnover during terminal transferase (Pimentel et al. 2022). Plasmids encoding RT proteins used in OTTR are available from AddGene: 2Bc-T MBP_BoMoC(ed)_6xH (Relay Pol) and 2Bc-T MBP_BoMoC(ed)_F753A_6xH (Tail Pol). Detailed OTTR protocols and reagent preparation instructions are included as Supplemental Information. To 40 ng of RNA in 9 µL of nuclease-free water, 2 µL of Buffer 1A (140 mM Tris-HCl pH 7.5, 1 M KCl, 14 mM DTT, 35% PEG-8000) and 1 µL of Buffer 1B (28 mM $MnCl_2$, 3.5 mM ddATP) was added and mixed well before 1 µL of 10 µM Tail Pol was added. The reaction was incubated for an hour and 30 minutes at 30 °C before 1 µL of Buffer 1C (3.5 mM ddGTP) was added. The reaction was well mixed once more before incubating for an additional 30 minutes at 30 °C. Unincorporated ddRTP molecules were inactivated by adding 1 µL of Buffer 2 (80 mM $MgCl_2$) and 1 µL of 1:1 Shrimp Alkaline Phosphatase (rSAP) (NEB):50% glycerol, mixing, incubating for

15 minutes at 37 °C. Then, 1 µL of Buffer 3 (100 mM EGTA), the reaction was mixed and incubated at 65 °C for 3 minutes. The reaction was immediately transferred to ice before 1 µL of Buffer 4A (10 mM MgCl2, 900 mM KCl, 40% PEG-6000, 4 mM dGTP, 0.8 mM dTTP, 0.8 mM dCTP, 0.04 mM dATP, 3 mM 2,6-diaminopurine-2′-deoxyribose-5′-triphosphate) and 1 µL of Buffer 4B (1.8 µM +1dY DNA/RNA adapter duplex primer, and either 3.6 µM 3′C adapter template or 3.6 µM 3′C unique molecular identifier (UMI) containing adapter template; See **Table 3.1**) was added and the reaction was vortexed. Then, 1 µL of 10 µM Relay Pol was added and the reaction was well mixed before incubating at 37 °C for 30 minutes followed by a 70 °C incubation for 5 minutes. To aid complete cDNA recovery, 0.5 µg/µL RNase A (Sigma, R6513) was added to the reaction and it was incubated at 50 °C for 5 minutes. 80 µL of cDNA extraction buffer (50 mM Tris pH 8, 20 mM EDTA, 0.2% SDS) was added to the reaction followed by 100 µL of phenol:chloroform:isoamyl alcohol (25:24:1) (PCI); the tube was vortexed and the aqueous phase was recovered. RNA was then precipitated by adding 10 uL of 7.5 M ammonium acetate and 300 uL of 100% ethanol and then incubating at -80 °C for 3 hours. For RNA templates > 40 nt, 0.5 U of Thermostable RNase H (NEB) was added to RNase A digestion, and Proteinase K was added to the cDNA extraction buffer and incubated at 50 °C for 15 minutes followed by denaturation at 95 °C for 5 minutes before PCI extraction.

Ligation-based ribosome profiling library generation.

The protocol from McGlincy and Ingolia 2017 was followed for ligation-based ribosome profiling. RNase I RPFs were dephosphorylated with T4 PNK before splitting samples for ligation-based and OTTR library generation. In this study, two pools of T4 PNK-treated RPFs were produced from a single batch of yeast cell lysate, but only one pool yielded a successful ligation-based library. Library generation failure, in this case, was because one RNA adapter ligation reaction was more inefficient than the other, leading to incongruent pooling following linker ligation.

cDNA size selection, elution, and quantification.

OTTR reaction products were resuspended in bromophenol blue formamide loading dye (xylene cyanol was excluded as it co-migrates with the ~75 nt primer dimer and interferes with Cy5 detection). The RNA oligonucleotide size markers used for RPF selection were carried through cDNA synthesis and used as positive controls and size markers for cDNA isolation. OTTR cDNAs were resolved by electrophoresis in 0.6X TBE 8% Urea-PAGE and visualized without staining using the fluorescence of the 5′ Cy5-dye. Excised urea-PAGE gel slices were crushed against the walls of a 1.5 mL tube using a sterile 1 mL pipette tip and cDNA was eluted by incubating at 70 °C for an hour in cDNA elution buffer containing 300 mM NaCl, 1 mM EDTA, and 10 mM Tris pH 8.0. cDNA was precipitated from the eluate using 3X volume of 100% ethanol followed by a 1 hour incubation at -80 C or snap-freezing in liquid nitrogen. The precipitate was

centrifuged for 15 minutes at 20,000 × g at 4 °C and the cDNA pellet was washed at least once with 70% ethanol and resuspended in 25 µL double distilled, nuclease-free water. We followed a previously described procedure to quantify cDNA by real-time PCR in order to determine the number of PCR cycles needed for library amplification (McGlincy and Ingolia 2017). We typically used 10 – 12 PCR cycles, and diluted some cDNA libraries to ensure all libraries in a cross-comparison cohort (e.g., all libraries in Figure **3.**3) were amplified with identical conditions. See indexing primers **Table 3.1**.

mirRICH small RNA enrichment from sucrose cushion pelleted ribosomes.

Ribosome pellets were resuspended in 60 µl polysome buffer, combined with 600 µL TRIzo,l and incubated for 10 minutes. The sample was then split in half, with 330 µL purified by Direct-Zol and 330 µl by mirRICH (Choi et al. 2018). Direct-Zol purified RNA was eluted in 100 µL, 10 µL 3 M sodium acetate pH 5.5 and 300 µL of 100% ethanol was added, and RNA was precipitated at -80 ºC for 3 hours. For mirRICH samples, we added chloroform, vortexed, centrifuged, and isolated the aqueous phase as described by the manufacturer. RNA was precipitated by adding 100% isopropanol and incubating on ice for 10 min. RNA was pelleted by centrifugation at 20,000 × g for 15 minutes at 4 °C, all liquid was removed, and the pellet was air dried for at least 2 hours. 10 µL of double-distilled water was added to the air-dried pellet and RNA was eluted for 10 minutes. The eluate was not mixed or vortexed during the elution period. 9 µL of the eluate liquid was removed, leaving the remaining pellet and ~1 µL behind, and the eluate was further purified by RNA Clean & Concentrator-5 and quantified by Nanodrop. OTTR libraries were made from ~100 ng of Direct-Zol purified RNA or ~40 ng of mirRICH-enriched RNA (~1% of the total RNA or ~10% of mirRICH RNA from the same amount of extract), and cDNA size selection was performed using OTTR cDNA generated from synthetic RNA input size markers.

Ligation-based and OTTR adapter trimming.

Ligation-based library reads were trimmed as described using Cutadapt (v3.2) (McGlincy and Ingolia 2017). A custom script was used to partition reads based on the ligated adapter's linker sequence, allowing for up to one mismatch. For the OTTR library in Figure **3.**1, the first base (usually a C or T) was removed from each read, and the DNA/RNA adapter duplex primer sequence (-a RGATCGGAAGAGCACACGTCTGAACTCCAGTCAC) was trimmed. For all libraries, any trimmed reads shorter than 20 or of poor quality (-q 10) were discarded. For OTTR libraries after Figure **3.**1, a UMI (CYNNNNN) was included. The UMI sequence immediately follows after the R1 sequence in the 3′C adapter template (see **Table 3.1**). When OTTR UMI was included, the following Cutadapt pipeline was used to trim the read and annotate the UMI and primer +1 base (i.e., +1T or +1C in this study) in the FASTQ ID:

cutadapt -m 2 -a GATCGGAAGAGCACACGTCTGAACTCCAGTCAC $tmpvar.fastq | cutadapt -u 7 --rename='{id} NTA={cut_prefix}' - | cutadapt -u -1 --rename='{id}_{comment}_primer={cut_suffix}' - | cutadapt -m 20 -q 10 –

Yeast reference sequences

Yeast rRNA sequences RDN18-1, RDN18-2, RDN25-1, RDN25-2, RDN37-1, RDN37-2, RDN5-1, RDN5-2, RDN5-3, RDN5-4, RDN5-5, RDN5-6, RDN58-1, RDN58-2, Q0020, and Q0158 from *Saccharomyces* Genome Database (SGD) were concatenated. The yeast tRNA sequences and reference were prepared as described (Pimentel et al. 2022). Non-coding RNA reference was made from the ncRNA, snoRNA, and snRNA sequences from SGD. Two different mRNA references were used in this paper. For analysis related to iχnos we used the previously described mRNA reference, except we extended the 5′ untranslated region (UTR) and 3′ UTR sequences to include up to 50 bases (Tunney et al. 2018). We also constructed an mRNA reference where the SGD mRNA CDS sequence was matched with the longest annotated 5′ UTR and 3′ UTR sequence using a custom script that combines the CDS coordinates with the annotated BED coordinates (Nagalakshmi et al. 2008, Yassour et al. 2009). In total, 5432 yeast mRNA sequences were constructed with either an annotated 5′ UTR and/or 3′ UTR sequence. 432 Ty-element mRNA sequences and 396 intron-containing, CDS-only, or frameshifting sequences were identified and isolated to a separate reference

Human reference sequences

Human rRNA sequences NR_023379.1, NR_146151.1, NR_146144.1, NR_145819.1, NR_146117.1, NR_003285.3, NR_003286.4, X12811.1, and NR_003287.4 from NCBI and ENST00000389680.2 and ENST00000387347.2 from Ensembl were concatenated. The human tRNA sequences and references were prepared as described (Holmes et al. 2022). ncRNA from Ensembl and lncRNA from Gencode were concatenated. The primary assembly of the human gDNA from NCBI, GRCh38, was used. 18,640 NCBI RefSeq MANE v0.95 transcripts were used, and the 5′ UTR, CDS, and 3′ UTR lengths were parsed from the GenBank entries using a custom script. The 13 mitochondrial mRNA sequences from Ensembl were also included in the mRNA reference, but typically excluded from analysis.

General ribosome profiling bioinformatics

We removed contaminating reads with a sequential analysis pipeline. Trimmed reads were removed by mapping to the rRNA reference (bowtie -v 3 -a --best --norc), the tRNA reference by tRAX (Holmes et al. 2022), the ncRNA reference (bowtie -v 3 -a --best --norc), and the mtDNA reference (bowtie -v 2 -a --best --norc). The remaining reads were then mapped to the mRNA references (bowtie -v 2 -m 200 -a --norc) and

sorted (samtools sort -n). Lastly, we sequentially mapped remaining reads to the gDNA reference (bowtie -v 2 -a --best)before mapping unaligned reads to a reference of size-selecting oligos and adapter oligos (bowtie -v 2 -a --best). To quantify gene-level CDS occupancy we used Bedtools intersect (v2.25.0) to exclude reads which aligned outside of the +15th to -10th codon of the CDS, counted reads using RSEM (rsem-calculate-expression --alignments --strandedness forward --seed-length 20 --sampling-for-bam), and then normalized and analyzed gene counts using DESeq2 (Li and Dewey. 2011, Love et al. 2014, Chothani et al. 2019). Two replicates were produced per condition, except for **Fig. 3.1B**'s which were a direct comparison of two library methods from a single RPF sample. DESeq2 normalized (median-of-ratios) gene-level counts for each replicated condition were averaged. For codon- or nucleotide-level plots, rescaled counts were used, which was the counts for each position in the transcript divided by the average number of reads per codon on that transcript.

Initiation, termination, and solitary codon occupancy profiles

Only genes with ≥ 50 nt of annotated 5′ UTR, ≥ 450 nt of annotated CDS, ≥ 50 nt of annotated 3′ UTR, and ≥ 1 reads per codon were considered for this analysis. Alignments were counted based on read length and either 5′ or 3′ position relative to the first base of a codon of interest (i.e., either start codon, stop codon, or solitary codon), and rescaled by average number of reads per codon for a given gene. Rescaled counts for each read length and relative position were summed and divided by the either the number of genes under consideration (i.e., for features that exist once per gene, like initiation and termination) or the number of windows under consideration (i.e., for solitary histidine codon loci across the translatome). To identify specific sites of true-disome accumulation, positions of trailing ribosome A-site codon positions were identified and the read lengths were binned (e.g., read lengths of 58, 59, and 60 nt were binned to 58 – 60 nt). Assuming that true disome RPFs are at least 58 nt, we determined the top five disome positions in the translatome where true-disomes occur by ranking the rescaled counts.

A-site occupancy comparisons

Only cytosolic genes with ≥ 50 nt of annotated 5′ UTR, ≥ 450 nt of annotated CDS, ≥ 50 nt of annotated 3′ UTR, and ≥ 1 reads per codon were considered for this analysis. Alignments outside of the first 15 and final 10 codons were ignored. A-site offsets were predetermined based on alignments to the initiation profile. A-site assignments were counted for each gene and rescaled by average number of reads per codon. The rescaled occupancy for each codon, excluding either initiation or termination codons, were averaged across the translatome, and the final A-site occupancy estimates were rescaled out of 61 before comparing.

iχnos full-model and leave-one-out models.

We extended the iχnos yeast transcriptome to include a total of 50 nt of both 5′ UTR and 3′ UTR sequences. After rRNA-, tRNA-, and ncRNA-mapping reads were excluded, Bowtie was used to map the reads against iχnos's yeast mRNA transcriptome (bowtie -v 2 --norc -a -m 250), and multi-mapping reads were reconciled by RSEM to use the use the highest-scoring multi-mapping read (rsem-calculate-expression --alignments --strandedness forward --seed-length 20 --sampling-for-bam) (Tunney et al. 2018, Li and Dewey. 2011).

A total of 1000 genes met the criteria to be studied by iχnos and were split into training (n=674) and testing (n=326) gene sets. iχnos was run with paralog-sensitive parameters on RPF alignments ranging from 25 to 29 nt in length were selected, and those which mapped to codons after the 20th but before the -20th codon were evaluated. Alignments were assigned to codons using A-site offset, and the codon-level counts for a given gene were normalized by the gene's average occupancy. From this, we used a previously described iχnos model of 13 codons from the -5th codon upstream to the +7th codon downstream (n5p7) of a given A-site. We first trained the n5p7-model five times for 45 epochs at a learning rate of 16 for each library and measured Pearson's correlation between predicted and actual occupancy for the testing set. We performed the same training regimen on 13 different model designs that only included 12 of the 13 codons, leaving one out (LOO) for each of the 13 different models and measured the average change in Pearson's correlations for each of the 13 LOO-models compared to the full-model.

Tables

Table 3.1: Oligonucleotides used in OTTR ribosome profiling of P1 nuclease RPFs

P1 nuclease 30 nt RNA marker	/5Phos/rNrGrArCrArGrArCrUrGrArCrUrArCrUrCrACrArCrGrArArCrArGrGrArN
P1 nuclease 40 nt RNA marker	/5Phos/rNrArGrUrCrGrUrCrUrCrArUrCrArGrGrUrCrUrCrUrCrArCrUrCrArCrUrArCrArCrArCrUrCrUrCrN
Cy5 +1T primer strand (DNA)	/5Cy5/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Cy5 +1C primer strand (DNA)	/5Cy5/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCC
Primer strand complement (RNA)	rGrArUrCrGrGrArArGrAmGmCmAmCmAmCmGmUmCmUmGmAmAmCmUmCmCmAmGmU/3SpC3/
3′C adapter template	/5Cy5/ACACTCTTTCCCTACACGACGCTCTTCCrGrArUrCrUrC
3′C UMI adapter template	/5Cy5/ACACTCTTTCCCTACACGACGCTCTTCCGATCTNrNrNrNrNrYrC
i5 indexing primer	AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGAC
i7 indexing primer	CAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTG

Notes: Oligonucleotides synthesized and used in this study for RNA size selection, input cDNA synthesis controls, OTTR primer-duplex construction, and either UMI-free or UMI-containing OTTR 3′C chimeric adapter templates. The OTTR oligonucleotides used in this study have been improved since the construction of these libraries. Several oligonucleotide modifications are used. /5Phos/ refers to a 5′ phosphate, /5Cy5/ refers to a 5′ Cy5, and /3SpC3/ refers to a 3′ three-carbon linker modification. Bases preceded by rare RNA nucleotides, and bases preceded by m are 2′-O-methyl RNA nucleotides. N bases are degenerate and Y bases are pyrimidines. The eight nucleotides of indexing i5 and i7 sequences are ablated from the sequence but their positions are denoted with [i5] and [i7], respectively.

Chapter 4: Additional considerations and future improvements to OTTR

A remaining challenge for OTTR is to limit production or improve removal of single-jump adapter-dimer cDNAs, which lack an insert sequence. For many applications, for example ribosome profiling (Ferguson et al. 2023), the presence of adapter-dimer cDNAs necessitates OTTR cDNA size selection by dPAGE. This can be a challenge, especially for laboratories that lack a suitable fluorescence imager or rely on commercially manufactured dPAGE gels, which in our experience fail to prevent reannealing of adapter oligonucleotides to OTTR cDNA during electrophoresis. While we found that changes to the 3′ddC adapter-template sequence could enhance (*e.g.,* 3′ddCYNNNNN) or suppress (*e.g.,* 3′ddCRNNNNN) adapter-dimer cDNA synthesis, we have not found a 3′ adapter-template sequence that suppresses adapter dimer formation to an extent that eliminates the utility of cDNA size selection by dPAGE. To date, we have maintained a 2:1:4 ratio of primer duplex:input template:adapter template and used these oligonucleotides at fixed concentrations suitable for up to 0.9 pmol of input RNA (Upton et al. 2021). For lower-input libraries, establishing a context-dependent strategy to reduce adapter dimer formation is worth considering, whether it be dilution of enzyme, adapter oligonucleotides, or both.

Approaches to limit adapter-dimer presence in sequencing libraries can also take advantage of molecular features that distinguish an adapter-dimer from OTTR cDNAs. One obvious difference is the fact that adapter-dimer cDNAs contain only sequences derived from the +1 T primer duplex and 3′ddC adapter template (**Table 2.3**). The proximity of the UMI to the adapter-dimer junction of +1 T primer duplex and 3′ddC adapter template complicates the design of antisense guides for targeted cleavage by programmable nucleases like *Thermus thermophilus* argonaute (Swarts et al. 2014). Another discriminating feature is that only OTTR cDNA duplexes contain a 3′ddA-labeled input template RNA. Conveniently, we have found that biotinylated deoxyadenosine triphosphate (dATP) or biotinylated ddATP can be used for 3′ labeling, and that input templates with a biotinylated 3′-nucleotide are readily captured for OTTR cDNA library synthesis (data not shown). Enrichment dependent on the inclusion of a 3′-labeled input template would both remove adapter-dimer cDNA and eliminate 3′ imprecision from over-capture (**Fig. 2.2A**).

OTTR is amenable to bespoke tailoring. To restrict input nucleic acid capture to be specific for particular 3′-end sequences (*e.g.,* 3′rG RNAs produced by RNase T1 cleavage), a custom-designed primer duplex can be used (for 3′rG RNAs, primer duplex with +1 C made from annealed +C.v3 and RNA.v3, **Table 2.6**). For input-template capture without an initial 3′-labeling step, a mixture of +1 primer duplexes can be used (**Table 2.6,** v3 +1 set). Because +1 G primer duplex with 3′ddC adapter template would efficiently generate adapter dimer, splitting the OTTR cDNA library synthesis reaction into successive, single template-jump reactions would be prudent. The first template-jump would synthesize cDNA across the input template using a BoMoC variant and/or reaction conditions that limit NTA. The second template-jump would be initiated by providing the 3′ddC adapter template in reaction conditions that support +1 G NTA.

From the perspective of cDNA purification, the use of a primer duplex with passenger-strand RNA modified by 2′-O-methyl groups (Upton et al. 2021) and a DNA-only, 3′ddC adapter template (developed in this study) prevent removal of these cDNA-complementary oligonucleotides by RNase. This allowed both adapter oligonucleotides (**Table 2.5,** c3t_ddCR5N; **Table 2.2**, RNA.v2) to re-anneal with cDNA during dPAGE-based size-selection, which increased adapter-dimer cDNA contamination of the gel region with the desired OTTR cDNAs (**Extended Data 2.3A**). One adjustment would be to modify the primer-duplex passenger strand to contain interspersed RNA sufficient for passenger-strand fragmentation by RNase H into subsequences that do not reanneal at ambient temperature (**Table 2.2**, RNA.v2.1, and **Table 2.6**, RNA.v3.1). A second adjustment would be to immobilize the 5′ end of the 3′ddC adapter template such that the cDNA duplex could be captured by streptavidin and the cDNA strand released from the 3′ddC adapter template by denaturation.

To date, we are unaware of sequence preferences that would prevent a primer duplex or adapter template of any sequence from being used in OTTR. The same reaction conditions would be used for any primer duplex sequence with 3′ +1 T and any adapter template sequence with 3′ddC. For example, we previously used longer Illumina adapters to enable sequencing-ready cDNA library synthesis without PCR (Upton et al. 2021). It will be of interest to test adapter sequences used in applications other than Illumina sequencing and to test different UMI lengths and/or sample-identifying barcodes to support single-cell applications (*e.g.,* **Table 2.5,** c3t_ddCR5N_barcoded).

It is also of interest to test a broader range of input templates for OTTR. Beyond ncRNA and mRNA-fragment templates, OTTR should be suitable for production of sequencing libraries from forensic samples, ancient DNA, cell-free DNA, and formalin-fixed paraffin-embedded biopsy material given the high tolerance of BoMoC for base-modified templates such as mature tRNAs (Gustafsson et al. 2022, d'Almeida et al. 2023, Manning et al. 2024) and the low input requirement demonstrated above. OTTR remains to be optimized for these sample types and for lengths of nucleic acid longer than tRNA. Both RNA and DNA oligonucleotides can be efficiently 3′ labeled by BoMoC (Upton et al. 2021), but optimal accessibility of the 3′ end of longer DNA or RNA templates may require single-stranded binding proteins or other additives to suppress any preferential input capture. Pre-denaturation of template structure may be advantageous before the 3′-labeling step and again before cDNA synthesis.

Finally, optimizations that facilitate automation and/or sub-microliter scale reactions are areas of focus to enable OTTR for single-cell library synthesis. A single human cell has an estimated ~1.3 fg of miRNA in 10 pg of total RNA, roughly ~115,000 molecules (Peltier and Latham 2008, Bissels et al. 2009). Assuming that a hypothetical single-cell miRNA library could be enriched to at least 10% miRNA reads and 90% non-miRNA, with OTTR reaction conditions used above, we extrapolate an input requirement of ~230 cells to have less than 10% of mappable reads produced from *E. coli* nucleic acids. Beyond miRNAs, the use of OTTR for single-cell ribosome profiling is equally worth considering. From 10 pg of total RNA in a human cell, assuming ~9 pg rRNA, if that rRNA was entirely assembled into translating ribosomes, we calculate a theoretical maximum of 43.7 fg mRNA RPFs from a single cell. Again assuming that a

hypothetical library could be enriched to contain 10% RPFs and 90% non-RPF small RNAs, we extrapolate an input requirement of ~7 cells to have less than 10% of mappable reads produced from *E. coli* nucleic acids. With optimization, single-cell surveys of miRNA and/or RPFs using OTTR could be a highly cost-efficient and information-rich alternative to current protocols for single-cell profiling of polyadenylated mRNAs (Vandereyken et al. 2023).

References

Androvic P, Benesova S, Rohlova E, Kubista M, Valihrach L. 2022. Small RNA-Sequencing for Analysis of Circulating miRNAs: Benchmark Study. *J Mol Diagn* 24: 386-394. doi: 10.1016/j.jmoldx.2021.12.006.

Archer SK, Shirokikh NE, Beilharz TH, Preiss T. 2016. Dynamics of ribosome scanning and recycling revealed by translation complex profiling. *Nature* 535: 570-574. doi: 10.1038/nature18647.

Balint E, Unk I. 2023. For the Better or for the Worse? The Effect of Manganese on the Activity of Eukaryotic DNA Polymerases. *Int J Mol Sci* 25: 363 doi: 10.3390/ijms25010363.

Bibiłło A, Eickbush TH. 2002. The reverse transcriptase of the R2 non-LTR retrotransposon: continuous synthesis of cDNA on non-continuous RNA templates. *J Mol Biol* 316: 459-73. doi: 10.1006/jmbi.2001.5369.

Bibiłło A, Eickbush TH. 2002. High processivity of the reverse transcriptase from a non-long terminal repeat retrotransposon. *J Biol Chem* 277: 34836-45. doi: 10.1074/jbc.M204345200.

Bibiłło A, Eickbush TH. 2004. End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. *J Biol Chem* 279: 14945-53. doi: 10.1074/jbc.M310450200.

Bissels U, Wild S, Tomiuk S, Holste A, Hafner M, Tuschl T, Bosio A. 2009. Absolute quantification of microRNAs by using a universal reference. *RNA* 15: 2375-84. doi: 10.1261/rna.1754109.

Bohlen, J., Fenzl, K., Kramer, G., Bukau, B. & Teleman, A. A. 2020. Selective 40S Footprinting Reveals Cap-Tethered Ribosome Scanning in Human Cells. *Mol Cell* 79: 561-574.e5. doi: 10.1016/j.molcel.2020.06.005.

Boskovic A, Bing XY, Kaymak E, Rando OJ. 2020. Control of noncoding RNA production and histone levels by a 5′ tRNA fragment. Genes Dev 34: 118-131. doi: 10.1101/gad.332783.119.

Boutabout M, Wilhelm M, Wilhelm FX. 2001. DNA synthesis fidelity by the reverse transcriptase of the yeast retrotransposon Ty1. *Nucleic Acids Res* 29: 2217-22. doi: 10.1093/nar/29.11.2217.

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH Jr. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A* 100: 5280-5. doi: 10.1073/pnas.0831042100.

Burke WD, Calalang CC, Eickbush TH. 1987. The site-specific ribosomal insertion element type II of Bombyx mori (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. *Mol Cell Biol* 7: 2221-30. doi: 10.1128/mcb.7.6.2221-2230.1987.

Chandramouly G, Liao S, Rusanov T, Borisonnik N, Calbert ML, Kent T, Sullivan-Reed K, Vekariya U, Kashkina E, Skorski T, et al. 2021. Polθ promotes the repair of 5′-DNA-protein crosslinks by microhomology-mediated end-joining. *Cell* Rep. 34: 108820. doi: 10.1016/j.celrep.2021.108820.

Choi C, Yoon S, Moon H, Bae YU, Kim CB, Diskul-Na-Ayudthaya P, Ngu TV, Munir J, Han J, Park SB, et al. 2018. mirRICH, a simple method to enrich the small RNA fraction from over-dried RNA pellets. *RNA Biol* 15: 763-772. doi: 10.1080/15476286.2018.1451723.

Chothani S, Adami E, Ouyang JF, Viswanathan S, Hubner N, Cook SA, Schafer S, Rackham OJL. 2019. deltaTE: Detection of Translationally Regulated Genes by Integrative Analysis of Ribo-seq and RNA-seq Data. *Curr Protoc Mol Biol* 129: e108. doi: 10.1002/cpmb.108.

Christensen S, Eickbush TH. 2004. Footprint of the retrotransposon R2Bm protein on its target site before and after cleavage. *J Mol Biol* 336: 1035-45. doi: 10.1016/j.jmb.2003.12.077.

Christensen SM, Eickbush TH. 2005. R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. *Mol Cell Biol* 25: 6617-28. doi: 10.1128/MCB.25.15.6617-6628.2005.

Christensen SM, Ye J, Eickbush TH. 2006. RNA from the 5′ end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site. *Proc Natl Acad Sci U S A* 103: 17602-7. doi: 10.1073/pnas.0605476103.

Coenen-Stass AML, Magen I, Brooks T, Ben-Dov IZ, Greensmith L, Hornstein E, Fratta P. 2018. Evaluation of methodologies for microRNA biomarker detection by next generation sequencing. *RNA Biol* 15: 1133-1145. doi: 10.1080/15476286.2018.1514236.

Cook GM, Brown K, Shang P, Li Y, Soday L, Dinan AM, Tumescheit C, Mockett APA, Fang Y, Firth AE, et al. 2022. Ribosome profiling of porcine reproductive and respiratory syndrome virus reveals novel features of viral gene expression. *Elife* 11: e75668. doi: 10.7554/eLife.75668.

d'Almeida GS, Casius A, Henderson JC, Knuesel S, Aphasizhev R, Aphasizheva I, Manning AC, Lowe TM, Alfonzo JD. 2023. tRNATyr has an unusually short half-life in Trypanosoma brucei. *RNA* 29: 1243-1254. Doi: 10.1261/rna.079674.123.

Deng P, Tan SQ, Yang QY, Fu L, Wu Y, Zhu HZ, Sun L, Bao Z, Lin Y, Zhang QC et al. 2023. Structural RNA components supervise the sequential DNA cleavage in R2 retrotransposon. *Cell* 186: 2865-2879.e20. doi: 10.1016/j.cell.2023.05.032.

Döring K, Ahmed N, Riemer T, Suresh HG, Vainshtein Y, Habich M, Riemer J, Mayer MP, O'Brien EP, Kramer G, et al. 2017. Profiling Ssb-Nascent Chain Interactions Reveals Principles of Hsp70-Assisted Folding. *Cell* 170: 298-311.e20. doi: 10.1016/j.cell.2017.06.038.

Eickbush TH, Eickbush DG. 2015. Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. *Microbiol Spectr* 3: MDNA3-0011-2014. doi: 10.1128/microbiolspec.MDNA3-0011-2014.

Ferguson L, Upton HE, Pimentel SC, Mok A, Lareau LF, Collins K, Ingolia NT. 2023. Streamlined and sensitive mono- and di-ribosome profiling in yeast and human cells. *Nat Methods* 20: 1704-1715. doi: 10.1038/s41592-023-02028-1.

Flynn RA, Pedram K, Malaker SA, Batista PJ, Smith BAH, Johnson AG, George BM, Majzoub K, Villalta PW, Carette JE, et al. 2021. Small RNAs are modified with N-glycans and displayed on the surface of living cells. *Cell* 184:3109-3124.e22. doi: 10.1016/j.cell.2021.04.023.

Fuchs RT, Sun Z, Zhuang F, Robb GB. 2015. Bias in ligation-based small RNA sequencing library construction is determined by adapter and RNA structure. *PLoS One* 10: e0126049. doi: 10.1371/journal.pone.0126049.

Fujimoto M, Kuninaka A, Yoshino H. 1974. Substrate Specificity of Nuclease P1. Agr Biol Chem 38: 1555-1561. doi: 10.1080/00021369.1974.10861384

Gaba A, Wang H, Fortune T, Qu X. 2021. Smart-ORF: a single-molecule method for accessing ribosome dynamics in both upstream and main open reading frames. *Nucleic Acids Res* 49: e26. doi: 10.1093/nar/gkaa1185.

Gao X, Wan J, Liu B, Ma M, Shen B, Qian SB. 2015. Quantitative profiling of initiating ribosomes in vivo. *Nat Methods* 12: 147-153. doi: 10.1038/nmeth.3208.

Garcés J, Wittek R. Reverse-transcriptase-associated RNaseH activity mediates template switching during reverse transcription in vitro. 1991. *Proc Biol Sci* 243: 235-239. doi: 10.1098/rspb.1991.0037.

Garshott DM, An H, Sundaramoorthy E, Leonard M, Vicary A, Harper JW, Bennett EJ. 2021. iRQC, a surveillance pathway for 40S ribosomal quality control during mRNA translation initiation. *Cell* Rep 36: 109642. doi: 10.1016/j.celrep.2021.109642.

Gerashchenko MV, Gladyshev VN. 2017. Ribonuclease selection for ribosome profiling. *Nucleic Acids Res* 45: e6. doi: 10.1093/nar/gkw822.

Giraldez MD, Spengler RM, Etheridge A, Godoy PM, Barczak AJ, Srinivasan S, De Hoff PL, Tanriverdi K, Courtright A, Lu S, et al. 2018. Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling. *Nat Biotechnol* 36: 746-757. doi: 10.1038/nbt.4183.

Gladyshev EA, Arkhipova IR. 2011. A widespread class of reverse transcriptase-related cellular genes. *Proc Natl Acad Sci U S A* 108: 20311-6. doi: 10.1073/pnas.1100266108.

Golinelli MP, Hughes SH. 2002. Nontemplated nucleotide addition by HIV-1 reverse transcriptase. Biochemistry 41: 5894-906. doi: 10.1021/bi0160415.

Gustafsson HT, Ferguon L, Galan C, Yu T, Upton HE, Kaymak E, Weng Z, Collins K, Rando OJ. 2022. Deep sequencing of yeast and mouse tRNAs and tRNA fragments using OTTR. bioRxiv doi: 10.1101/2022.02.04.479139.

Guydosh NR, Green R. 2014. Dom34 rescues ribosomes in 3′ untranslated regions. *Cell* 156: 950-62. doi: 10.1016/j.cell.2014.02.006.

Han Y, Gao X, Liu B, Wan J, Zhang X, Qian SB. 2014. Ribosome profiling reveals sequence-independent post-initiation pausing as a signature of translation. *Cell Res* 24: 842-851. doi: 10.1038/cr.2014.74.

Hem JD. 1963. Chemical equilibria and rates of manganese oxidation. U.S. G.P.O. 1667-A: 1-64. doi: 10.3133/wsp1667A.

Herbert ZT, Thimmapuram J, Xie S, Kershner JP, Kolling FW, Ringelberg CS, LeClerc A, Alekseyev YO, Fan J, Podnar JW, et al. 2020. Multisite Evaluation of Next-Generation Methods for Small RNA Quantification. *J Biomol Tech* 31: 47-56. doi: 10.7171/jbt.20-3102-001.

Hinnebusch AG. 2005. Translational regulation of GCN4 and the general amino acid control of yeast. *Annu Rev Microbiol* 59: 407-450. doi: 10.1146/annurev.micro.59.031805.

Holmes AD, Howard JM, Chan PC, Lowe TM. 2022. tRNA Analysis of eXpression (tRAX): A tool for integrating analysis of tRNAs, tRNA-derived small RNAs, and tRNA modifications bioRxiv doi: 10.1101/2022.07.02.498565.

Hornstein N, Torres D, Das Sharma S, Tang G, Canoll P, Sims PA. 2016. Ligation-free ribosome profiling of cell type-specific translation in the brain. *Genome Biol* 17: 149. doi: 10.1186/s13059-016-1005-1.

Hücker SM, Fehlmann T, Werno C, Weidele K, Lüke F, Schlenska-Lange A, Klein CA, Keller A, Kirsch S. 2021. Single-cell microRNA sequencing method comparison and application to cell lines and circulating lung tumor cells. *Nat Commun* 12: 4316. doi: 10.1038/s41467-021-24611-w.

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324: 218-223. doi: 10.1126/science.1168978.

Irigoyen N, Firth AE, Jones JD, Chung BY, Siddell SG, Brierley I. 2016. High-Resolution Analysis of Coronavirus Gene Expression by RNA Sequencing and Ribosome Profiling. *PLoS Pathog* 12: e1005473. doi: 10.1371/journal.ppat.1005473.

Irigoyen N, Dinan AM, Brierley I, Firth AE. 2018. Ribosome profiling of the retrovirus murine leukemia virus. *Retrovirology* 15: 10 doi: 10.1186/s12977-018-0394-5.

Jamburuthugoda VK, Eickbush TH. 2014. Identification of RNA binding motifs in the R2 retrotransposon-encoded reverse transcriptase. *Nucleic Acids Res* 42: 8405-15. doi: 10.1093/nar/gku514.

Jan CH, Williams CC, Weissman JS. 2014. Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. *Science* 346: 1257521. doi: 10.1126/science.1257521.

Jiang J, Miracco EJ, Hong K, Eckert B, Chan H, Cash DD, Min B, Zhou ZH, Collins K, Feigon J. 2013. The architecture of Tetrahymena telomerase holoenzyme. *Nature* 496: 187-92. doi: 10.1038/nature12062.

Kaikkonen MU, Lam MT, Glass CK. 2011. Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovasc Res* 90: 430-40. doi: 10.1093/cvr/cvr097.

Kainuma M, Hershey JW. 2001. Depletion and deletion analyses of eukaryotic translation initiation factor 1A in Saccharomyces cerevisiae. *Biochimie* 83: 505-514. doi: 10.1016/s0300-9084(01)01279-2.

Kent T, Mateos-Gomez PA, Sfeir A, Pomerantz RT. 2016. Polymerase θ is a robust terminal transferase that oscillates between three different mechanisms during end-joining. *Elife* 5: e13740. doi: 10.7554/eLife.13740.

Kovaľ T, Østergaard LH, Lehmbeck J, Nørgaard A, Lipovová P, Dušková J, Skálová T, Trundová M, Kolenko P, Fejfarová K, et al. 2016. Structural and Catalytic Properties of S1 Nuclease from Aspergillus oryzae Responsible for Substrate Recognition, Cleavage, Non-Specificity, and Inhibition. *PLoS One* 11: e0168832. doi: 10.1371/journal.pone.0168832.

Kurzynska-Kokorniak A, Jamburuthugoda VK, Bibiłło A, Eickbush TH. 2007. DNA-directed DNA polymerase and strand displacement activity of the reverse transcriptase encoded by the R2 retrotransposon. *J Mol Biol* 374: 322-33. doi: 10.1016/j.jmb.2007.09.047.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, Fitzhugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860-921. doi: 10.1038/35057062.

Lareau LF, Hite DH, Hogan GJ, Brown PO. 2014. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *Elife* 3: e01257. doi: 10.7554/eLife.01257.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323. doi: 10.1186/1471-2105-12-323.

Li Q, Yang H, Stroup EK, Wang H, Ji Z. 2022. Low-input RNase footprinting for simultaneous quantification of cytosolic and mitochondrial translation. *Genome Res* 32: 545-557. doi: 10.1101/gr.276139.121.

Li Z, Ferguson L, Deol KK, Roberts MA, Magtanong L, Hendricks JM, Mousa GA, Kilinc S, Schaefer K, Wells JA, et al. 2022. Ribosome stalling during selenoprotein translation exposes a ferroptosis vulnerability. *Nat Chem Biol* 18: 751-761. doi: 10.1038/s41589-022-01033-3.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15: 550. doi: 10.1186/s13059-014-0550-8.

Lu C, Meyers BC, Green PJ. 2007. Construction of small RNA cDNA libraries for deep sequencing. *Methods* 43: 110-7. doi: 10.1016/j.ymeth.2007.05.002.

Luan DD, Eickbush TH. 1995. RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. *Mol Cell Biol* 15: 3882-91. doi: 10.1128/MCB.15.7.3882.

Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* 72: 595-605. doi: 10.1016/0092-8674(93)90078-5.

Lulla V, Dinan AM, Hosmillo M, Chaudhry Y, Sherry L, Irigoyen N, Nayak KM, Stonehouse NJ, Zilbauer M, Goodfellow I, et al. 2019. An upstream protein-coding region in enteroviruses modulates virus infection in gut epithelial cells. *Nat Microbiol* 4: 280-292. doi: 10.1038/s41564-018-0297-1.

Manning AC, Bashir MM, Jimenez AR, Upton HE, Collins K, Lowe TM, Tucker JM. 2024. Gammaherpesvirus infection alters transfer RNA splicing and triggers t RNA cleavage. bioRxiv doi: 10.1101/2024.02.16.580780.

Marquet R, Isel C, Ehresmann C, Ehresmann B. 1995. tRNAs as primer of reverse transcriptases. *Biochimie* 77: 113-24. doi: 10.1016/0300-9084(96)88114-4.

Martin G, Keller W. 1998. Tailing and 3′-end labeling of RNA with yeast poly(A) polymerase and various nucleotides. *RNA* 4: 226-30.

McGlincy NJ, Ingolia NT. 2017. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* 126: 112-129. doi: 10.1016/j.ymeth.2017.05.028.

Mestre-Fos S, Ferguson L, Trinidad M, Ingolia NT, Cate JHD. 2023. eIF3 engages with 3′-UTR termini of highly translated mRNAs in neural progenitor cells. bioRxiv doi: 10.1101/2023.11.11.566681.

Meydan S, Guydosh NR. 2020. Disome and Trisome Profiling Reveal Genome-wide Targets of Ribosome Quality Control. *Mol Cell* 79: 588-602.e6. doi: 10.1016/j.molcel.2020.06.010.

Muller R, Meacham ZA, Ferguson L, Ingolia NT. 2020. CiBER-seq dissects genetic networks by quantitative CRISPRi profiling of expression phenotypes. *Science* 370: eabb9662. doi: 10.1126/science.abb9662.

Muthukumar S, Li CT, Liu RJ, Bellodi C. 2024. Roles and regulation of tRNA-derived small RNAs in animals. *Nat Rev Mol Cell* Biol. doi: 10.1038/s41580-023-00690-z.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320: 1344-1349. doi: 10.1126/science.1158441.

Naik AK and Raghavan SC. 2008. P1 nuclease cleavage is dependent on length of the mismatches in DNA. *DNA Repair* 7: 1384-1391. doi: 10.1016/j.dnarep.2008.04.010.

Napthine S, Ling R, Finch LK, Jones JD, Bell S, Brierley I, Firth AE. 2017. Protein-directed ribosomal frameshifting temporally regulates gene expression. *Nat Commun* 8: 15582. doi: 10.1038/ncomms15582.

Nguyen THD, Tam J, Wu RA, Greber BJ, Toso D, Nogales E, Collins K. 2018. Cryo-EM structure of substrate-bound human telomerase holoenzyme. *Nature* 557: 190-195. doi: 10.1038/s41586-018-0062-x.

O'Brien K, Breyne K, Ughetto S, Laurent LC, Breakefield XO. 2020. *RNA* delivery by extracellular vesicles in mammalian cells and its applications. *Nat Rev Mol Cell Biol* 21: 585–606. doi: 10.1038/s41580-020-0251-y.

Oh E, Becker AH, Sandikci A, Huber D, Chaba R, Gloge F, Nichols RJ, Typas A, Gross CA, Kramer G, et al. 2011. Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell* 147: 1295-1308. doi: 10.1016/j.cell.2011.10.044.

Ohtsubo Y, Nagata Y, Tsuda M. 2017. Efficient N-tailing of blunt DNA ends by Moloney murine leukemia virus reverse transcriptase. *Sci Rep* 7: 41769. doi: 10.1038/srep41769.

Park SK, Mohr G, Yao J, Russell R, Lambowitz AM. 2022. Group II intron-like reverse transcriptases function in double-strand break repair. *Cell* 185: 3671-3688.e23. doi: 10.1016/j.cell.2022.08.014.

Pelletier H, Sawaya MR, Wolfle W, Wilson SH, Kraut J. 1996. A structural basis for metal ion mutagenicity and nucleotide selectivity in human DNA polymerase beta. *Biochemistry* 35: 12762-77. doi: 10.1021/bi9529566.

Peltier HJ, Latham GJ. 2008. Normalization of microRNA expression levels in quantitative RT-PCR assays: Identification of suitable reference RNA targets in normal and cancerous human solid tissues. *RNA* 14: 844-52. doi: 10.1261/rna.939908.

Pimentel SC, Upton HE, Collins K. 2022. Separable structural requirements for cDNA synthesis, nontemplated extension, and template jumping by a non-LTR retroelement reverse transcriptase. *J Biol Chem* 298: 101624. doi: 10.1016/j.jbc.2022.101624.

Rapoport TA. 2007. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature* 450: 663-669. doi: 10.1038/nature06384.

Rodríguez-Vargas A, Collins K. 2024. Distinct and overlapping RNA determinants for binding and target-primed reverse transcription by Bombyx mori R2 retrotransposon protein. *Nucleic Acids Res* Epub ahead of print. doi: 10.1093/nar/gkae194.

Romier C, Dominguez R, Lahm A, Dahl O, Suck D. 1998. Recognition of single-stranded DNA by nuclease P1: high resolution crystal structures of complexes with substrate analogs. *Proteins* 32: 414-424. doi: 10.1002/(SICI)1097-0134(19980901)32:4<414::AID-PROT2>3.0.CO;2-G

Shi J, Zhou T, Chen Q. 2022. Exploring the expanding universe of small RNAs. *Nat Cell Biol* 24: 415–423. doi: 10.1038/s41556-022-00880-5

Shore S, Henderson JM, Lebedev A, Salcedo MP, Zon G, McCaffrey AP, Paul N, Hogrefe RI. 2019. Small RNA Library Preparation Method for Next-Generation Sequencing Using Chemical Modifications to Prevent Adapter Dimer Formation. *PLoS One* 11: e0167009. doi: 10.1371/journal.pone.0167009.

Shurtleff MJ, Itzhak DN, Hussmann JA, Schirle Oakdale NT, Costa EA, Jonikas M, Weibezahn J, Popova KD, Jan CH, et al. 2018. The ER membrane protein complex interacts cotranslationally to enable biogenesis of multipass membrane proteins. *Elife* 7: e37018. doi: 10.7554/eLife.37018.

Sonenberg N, Hinnebusch AG. 2009. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* 136: 731-745. doi: 10.1016/j.cell.2009.01.042.

Stein KC, Kriel A, and Frydman J. 2019. Nascent Polypeptide Domain Topology and Elongation Rate Direct the Cotranslational Hierarchy of Hsp70 and TRiC/CCT. *Mol Cell* 75: 1117-1130.e5. doi: 10.1016/j.molcel.2019.06.036.

Stewart H, Brown K, Dinan AM, Irigoyen N, Snijder EJ, Firth AE. 2018. Transcriptional and Translational Landscape of Equine Torovirus. *J Virol* 92: e00589-18 doi: 10.1128/JVI.00589-18.

Swarts DC, Jore MM, Westra ER, Zhu Y, Janssen JH, Snijders AP, Wang Y, Patel DJ, Berenguer J, Brouns SJJ, et al. 2014 DNA-guided DNA interference by a prokaryotic Argonaute. *Nature* 507: 258-261 doi: 10.1038/nature12971.

Thawani A, Ariza AJF, Nogales E, Collins K. 2024. Template and target-site recognition by human LINE-1 in retrotransposition. *Nature* 626: 186-193. doi: 10.1038/s41586-023-06933-5.

Tosar JP, Castellano M, Costa B, Cayota A. 2024. Small RNA structural biochemistry in a post-sequencing era. *Nat Protoc* 19: 595-602. doi: 10.1038/s41596-023-00936-2

Tunney R, McGlincy NJ, Graham ME, Naddaf N, Pachter L, Lareau LF. 2018. Accurate design of translational output by a neural network model of ribosome distribution. *Nat Struct Mol Biol* 25: 577-582. doi: 10.1038/s41594-018-0080-2.

Upton HE, Ferguson L, Temoche-Diaz MM, Liu XM, Pimentel SC, Ingolia NT, Schekman R, Collins K. 2021. Low-bias ncRNA libraries using ordered two-template relay: Serial template jumping by a modified retroelement reverse transcriptase. *Proc Natl Acad Sci U S A* 118: e2107900118. doi: 10.1073/pnas.2107900118.

van Niel G, Carter DRF, Clayton A, Lambert DW, Raposo G, Vader P. 2022. Challenges and directions in studying cell-cell communication by extracellular vesicles. *Nat Rev Mol Cell Biol* 23: 369-382. doi: 10.1038/s41580-022-00460-3.

Vandereyken K, Sifrim A, Thienpont B, Voet T. 2023. *Methods* and applications for single-cell and spatial multi-omics. *Nat Rev Genet* 24: 494–515. doi: 10.1038/s41576-023-00580-2.

VanInsberghe M, van den Berg J, Andersson-Rolf A, Clevers H, & van Oudenaarden A. 2021. Single-cell Ribo-seq reveals cell cycle-dependent translational pausing. *Nature* 597: 561-565. doi: 10.1038/s41586-021-03887-4.

Vashishtha AK, Wang J, Konigsberg WH. 2016. Different Divalent Cations Alter the Kinetics and Fidelity of DNA Polymerases. *J Biol Chem* 291: 20869-20875. doi: 10.1074/jbc.R116.742494.

Volbeda A, Lahm A, Sakiyama F, Suck D. 1991. Crystal structure of Penicillium citrinum P1 nuclease at 2.8 A resolution. *EMBO J* 10: 1607-1618. doi: 10.1002/j.1460-2075.1991.tb07683.x.

Wagner S, Herrmannová A, Hronová V, Gunišová S, Sen ND, Hannan RD, Hinnebusch AG, Shirokikh NE, Preiss T, Valášek LS. 2020. Selective Translation Complex Profiling Reveals Staged Initiation and Co-translational Assembly of Initiation Factor Complexes. *Mol Cell* 79: 546-560.e7. doi: 10.1016/j.molcel.2020.06.004.

Wang N, Zheng J, Chen Z, Liu Y, Dura B, Kwak M, Xavier-Ferrucio J, Lu YC, Zhang M, Roden C, et al. 2019. Single-cell microRNA-mRNA co-sequencing reveals non-genetic heterogeneity and mechanisms of microRNA regulation. 2019. *Nat Commun* 10: 95. doi: 10.1038/s41467-018-07981-6.

Wei J, Wu C, Sachs MS. 2012. The arginine attenuator peptide interferes with the ribosome peptidyl transferase center. *Mol Cell Biol* 32: 2396-2406. doi: 10.1128/MCB.00136-12.

Wilkinson ME, Frangieh CJ, Macrae RK, Zhang F. 2023. Structure of the R2 non-LTR retrotransposon initiating target-primed reverse transcription. *Science* 380: 301-308. doi: 10.1126/science.adg7883.

Williams CC, Jan, C. H. & Weissman, J. S. 2014. Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling. *Science* 346: 748-751. doi: 10.1126/science.1257522.

Wu CC, Zinshteyn B, Wehner KA, Green R. 2019. High-Resolution Ribosome Profiling Defines Discrete Ribosome Elongation States and Translational Regulation during Cellular Stress. *Mol Cell* 73: 959-970.e5. doi: 10.1016/j.molcel.2018.12.009.

Wu CC, Peterson A, Zinshteyn B, Regot S, Green R. 2020. Ribosome Collisions Trigger General Stress Responses to Regulate Cell Fate. *Cell* 182: 404-416. doi: 10.1016/j.cell.2020.06.006.

Wulf MG, Maguire S, Humbert P, Dai N, Bei Y, Nichols NM, Corrêa IR Jr, Guan S. 2019. Non-templated addition and template switching by Moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each other. *J Biol Chem* 294: 18220-18231. doi: 10.1074/jbc.RA119.010676.

Xiong Z, Xu K, Lin Z, Kong F, Wang Q, Quan Y, Sha QQ, Li F, Zou Z, Liu L, et al. 2022. Ultrasensitive Ribo-seq reveals translational landscapes during

mammalian oocyte-to-embryo transition and pre-implantation development. *Nat Cell Biol* 24: 968-980. doi: 10.1038/s41556-022-00928-6.

Xu H, Yao J, Wu DC, Lambowitz AM. 2019. Improved TGIRT-seq methods for comprehensive transcriptome profiling with decreased adapter dimer formation and bias correction. *Sci Rep* 9: 7953. doi: 10.1038/s41598-019-44457-z.

Yan LL, Zaher HS. 2021. Ribosome quality control antagonizes the activation of the integrated stress response on colliding ribosomes. *Mol Cell* 81: 614-628. doi: 10.1016/j.molcel.2020.11.033.

Yassour M, Kaplan T, Fraser HB, Levin JZ, Pfiffner J, Adiconis X, Schroth G, Luo S, Khrebtukova I, Gnirke A, et al. 2009. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc Natl Acad Sci U S A*. 106: 3264-3269. doi: 10.1073/pnas.0812841106.

Yehudai-Resheff S, Schuster G. 2000. Characterization of the E. coli poly(A) polymerase: nucleotide specificity, RNA-binding affinities and RNA structure dependence. *Nucleic Acids Res* 28: 1139-44. doi: 10.1093/nar/28.5.1139.

Zhang, C., Wang, M., Li, Y., & Zhang, Y. 2022. Profiling and functional characterization of maternal mRNA translation during mouse maternal-to-zygotic transition. *Sci Adv* 8: eabj3967. doi: 10.1126/sciadv.abj3967.

Zhang X, Van Treeck B, Horton CA, McIntyre JJR, Palm SM, Shumate JL, Collins K. 2024. Harnessing eukaryotic retroelement proteins for transgene insertion into human safe-harbor loci. *Nat Biotechnol* Epub ahead of print. doi: 10.1038/s41587-024-02137-y.

Zhao C, Liu F, Pyle AM. 2018. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* 24: 183-195. doi: 10.1261/rna.063479.117.

Zhao T, Chen YM, Li Y, Wang J, Chen S, Gao N, Qian W. 2021. Disome-seq reveals widespread ribosome collisions that promote cotranslational protein folding. *Genome Biol* 22: 16. doi: 10.1186/s13059-020-02256-0.

Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD. 2001. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* 30: 892-7. doi: 10.2144/01304pf02.